

The Phd thesis “BioSilicoSystems - A Multipronged approach towards analysis and representation of biological data” was submitted by me to the faculty of technology at Bielefeld University, Germany and later was approved at the faculty conference held in December 2009.

The submitted work is advised revision. I will upload the revised version in future.

Regards,
Sridhar Hariharaputran
November 2010.

Universität Bielefeld | Postfach 10 01 31 | 33501 Bielefeld

An den
Dekan der Technischen Fakultät
Prof. Dr. T. Noll

- im Hause -

Prof. Dr. Ralf Hofestädt

Raum: D5-106
Durchwahl: 0521.106-5283
Sekretariat: 0521.106-6885
Fax: 0521.106-6488
hofestae@techfak.uni-bielefeld.de
www.techfak.uni-bielefeld.de/ags/bi/

Bielefeld, den 08.12.2009

Seite 1 von 1

Bescheinigung

Sehr geehrte Damen und Herren,

hiermit bescheinige ich, dass ich die Dissertation von Herrn Sridhar Hariharaputran begutachten werde.

Mit freundlichen Grüßen



Prof. Dr. R. Hofestädt

Universität Bielefeld | Postfach 10 01 31 | 33501 Bielefeld

Herrn
Sridhar Hariharaputran
AG Bioinformatik/Medizininformatik

– im Hause –

Prof. Dr. Thomas Noll

Raum: C3-142
Tel.: 0521.106-00
DW: 0521.106-3462
Fax: 0521.106-6468
anke@techfak.uni-bielefeld.de
www.techfak.uni-bielefeld.de

Bielefeld, 17. Dezember 2009

Seite 1 von 1

Eröffnung des Promotionsverfahrens

Ihr Antrag vom 10.12.2009


Sehr geehrter Herr Hariharaputran,

ich teile Ihnen mit, dass die 172. Fakultätskonferenz am 16.12.2009 die Eröffnung Ihres Promotionsvorhabens beschlossen hat.

Als Gutachter wurden Herr Prof. Dr. Ralf Hofestädt und Herr Prof. Dr. Robert Giegerich bestellt. In den Prüfungsausschuss wurden außerdem Herr Prof. Dr. Jens Stoye und Frau Dr. Susanne Schneiker-Bekel gewählt.

Für das weitere Verfahren wünsche ich Ihnen alles Gute.

Mit freundlichen Grüßen



Prof. Dr. Thomas Noll

BioSilicoSystems

A Multipronged Approach Towards Analysis and Representation of Biological Data

Dissertation

submitted for the degree of
Doktor der Naturwissenschaften
(Dr. rer. nat.)

by

Sridhar Hariharaputran

Bielefeld University
Faculty of Technology
Germany

December 2009

Sridhar Hariharaputran
Bielefeld University
Faculty of Technology
Bioinformatics / Medical Informatics Department &
Graduate College Bioinformatics (GK 635)
Email: sharihar@techfak.uni-bielefeld.de

To my family

Acknowledgements

First and foremost I would like to thank *Prof. Ralf Hofestädt* my supervisor for giving me the wonderful opportunity to pursue my PhD work under his able guidance and for inspiring me to work on these projects. I am grateful to him for giving me the freedom to work on multidisciplinary projects and for his valuable suggestions and discussions which enabled me to tackle the problems and prepared me for future challenges. It is my great privilege to be associated with him and I would like to express my sincere gratitude.

I am grateful to *Prof. Robert Giegerich* speaker of the Graduate College Bioinformatics who has been encouraging me all these years. I owe my gratitude and I sincerely thank him for his motivation and support.

I take this opportunity to thank sincerely *Prof. Karl Josef Dietz*, *Prof. Karsten Niehaus* and his colleagues *Dr. Frank-Jörg Vorhölter*, *Dr. Thomas Patschkowski* for giving me their time and for their valuable discussions and comments which helped me in my projects. I would like to thank sincerely *Prof. Thomas Dierks*, *Prof. Hermann Ragg*, *Prof. Jens Stoye*, *Prof. Sven Rahmann* for giving me the opportunity to discuss my works and also to present my ideas.

I sincerely thank *Prof. Patrizio Arrigo* and his colleagues, *Prof. Francisco Azuaje* & his colleagues for the wonderful interactions, encouragement and support while working in the project and also for helping me with the data. I thank sincerely *Prof. Chanchal K Mitra*, *Prof. Falk Schreiber*, *Prof. Nils P Wilassen*, for their valuable comments, discussions, suggestions and encouragement. I would like to thank sincerely *Prof. Nikolay A Kolchanov* & his colleagues for giving me the opportunity to work with their software and also for their valuable suggestions and encouragement.

I am grateful to my previous investigators *Prof. Upinder S Bhalla* and *Prof. Nagasuma R Chandra* for giving me the opportunity to work in different areas which helped me in my PhD work and also for their encouragement and guidance. I take this opportunity to thank *Prof. Sekar K*, *Prof. RamaKumar S*, *Prof. Saraswathi Vishveshwara*, *Prof. Srinivasan N*, *Prof. Suguna K*, *Prof. Murthy MRN*, *Prof. Balakrishnan N* for their valuable comments, discussions and for their encouragement and support.

I am grateful to *Prof. Judith-Klein Seetharaman*, *Prof. Peter Bramley* for their encouragement and support. I take this opportunity to thank sincerely *Prof. Baldo Oliva Miguel*, *Prof. Christopher Reynolds* for their support and encouragement. I would like to thank *Prof. Arne Elofsson* for his encouragement, for the valuable suggestions and discussions.

I take this opportunity to thank all my present and past colleagues *David Braun*, *Prof. Ming Chen*, *Dr. Andreas Freier*, *Dr. Jacob Köhler*, *Benjamin Kormier*, *Dr. Dieter Lorenz*, *Hang Mao-Lee*, *Mark Niemann*, *Alijona Oprawachat*, *Benjamin Prins*, *Alexander Ruegg*, *Björn Sommer*, *Arben Soshi*, *Dr. Thoralf Töpel*. I am thankful to them for their immense and timely help and for their valuable comments, discussions and support. It was wonderful working them all these years and sharing memorable moments together.

I would like to thank sincerely *Ms. Sabine Klussmann* secretary of the department for her warmth and support and helping me all these years. Also, I would like to thank *Ms. Britta Quisbrok* secretary at the Graduate College Bioinformatics for her kind help and support during these years. I would also like to thank *Mr. Klaus Kulitza* and *Ms. Tanja Möller* for their timely help.

I would like to thank students *Björn Brockschmidt, Corina Fietz, Antonia Gross, Alexander Medina, Timm Oberwahrenbrock, Nils Oppermann, Sarah Spangardt & other project students*. It was a nice experience guiding and working with them in different areas and projects which motivated me to a great extent.

I also take this opportunity to thank all my friends at the Graduate College Bioinformatics (GK 635) and in the Graduate School (GS). It was wonderful discussing with them various topics and sharing nice moments. I take this opportunity to thank the staff in different departments within and outside the university who helped me at various times making my things easier and my stay memorable.

I would like to thank all my friends from India and other places whom I met in Germany for their encouragement and support. And I would like thank all my friends back home in India and elsewhere for their love and affection.

My sincere thanks to all others who are not explicitly mentioned anywhere in this dissertation for their encouragement and support.

I would like thank the funding agencies the DFG - Deutsche Forschungsgemeinschaft (German Research Foundation) Graduate College Bioinformatics (GK635), the Bioinformatics/Medical Informatics Department, EU Project CardioWorkBench - Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analyses and the University of Bielefeld for encouraging my work and for their financial support.

I am thankful to *Dr. Michael Hucka* and the organisers of The Ninth Workshop on Software Platforms for Systems Biology and to *Prof. Joseph Nadeau* and the organisers of Aegean Conference of 5th International Conference on Pathways, Networks and Systems for their fellowships and also for giving me the opportunity to attend and present my work.

All this would have not been possible without my family. Everything I owe to my parents, my brother and my sister-in-law. Their love, encouragement and consistent support in all these years provided me great strength and my heartfelt thanks to them.

In Vain Without God

With best regards
Sridhar Hariharaputran
Bielefeld

Summary

The rising field of integrative bioinformatics provides the vital methods to integrate, manage and also to analyze the diverse data and allows gaining new and deeper insights and a clear understanding of the intricate biological systems. The difficulty is not only to facilitate the study of heterogeneous data within the biological context, but it also more fundamental, how to represent and make the available knowledge accessible. Moreover, adding valuable information and functions that persuade the user to discover the interesting relations hidden within the data is, in itself, a great challenge. Also, the cumulative information can provide greater biological insight than is possible with individual information sources. Furthermore, the rapidly growing number of databases and data types poses the challenge of integrating the heterogeneous data types, especially in biology. This rapid increase in the volume and number of data resources drive for providing polymorphic views of the same data and often overlap in multiple resources.

In this thesis a multi-pronged approach is proposed that deals with various methods for the analysis and representation of the diverse biological data which are present in different data sources. This is an effort to explain and emphasize on different concepts which are developed for the analysis of molecular data and also to explain its biological significance. The hypotheses proposed are in context with various other results and findings published in the past. The approach demonstrated also explains different ways to integrate the molecular data from various sources along with the need for a comprehensive understanding and clear projection of the concept or the algorithm and its results, but with simple means and methods. The multifarious approach proposed in this work comprises of different tools or methods spanning significant areas of bioinformatics research such as data integration, data visualization, biological network construction / reconstruction and alignment of biological pathways. Each tool deals with a unique approach to utilize the molecular data for different areas of biological research and is built based on the kernel of the thesis. Furthermore these methods are combined with graphical representation that make things simple and comprehensible and also helps to understand with ease the underlying biological complexity. Moreover the human eye is often used to and it is more comfortable with the visual representation of the facts.

Publications

Following list show the publications from the PhD work

- Chen M, **Hariharaputran S**, Hofestädt R, Kormeier B and Spangardt S (2009). Petri Net Models for the Semi-automatic Construction of Large Scale Biological Networks. *Natural Computing*. DOI 10.1007/s11047-009-9151-y.
- **Hariharaputran S**, Hofestädt R, Kormeier B, Spangardt S (2008). Modelling and Visualisation of Pathways using Petri Nets. *In Proceedings of the 5th International Symposium on Integrative Bioinformatics, Wittenberg, Germany, August 20 - 22*. (Poster No.70)
- **Hariharaputran S**, Hofestädt R, Kormeier B, Spangardt S (2008). MOdeling and VISualisation of Pathways using Petri Nets. *In Proceedings of the Sixth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008), Novosibirsk, Russia, 22 -28 June*. (Poster B5)
- **Hariharaputran S***, Töpel T, Oberwahrenbrock T and Hofestädt R (2008). Alignment of Linear Biochemical Pathways Using Protein Structural Classification. *Nature Precedings* <http://dx.doi.org/10.1038/npre.2008.1943.1>.
- **Hariharaputran S***, Töpel T, Brockschmidt B and Hofestädt R (2008). VINEdb: a data warehouse for integration and interactive exploration of life science data. *Integrative Bioinformatics Yearbook 2007. Shaker Verlag Aachen 2008, 63-74*.
- **Hariharaputran S***, Töpel T, Brockschmidt B and Hofestädt R (2007). VINEdb: a data warehouse for integration and interactive exploration of life science data. *In Proceedings of the 4th Integrative Bioinformatics Workshop (IB 07), Ghent, Belgium, September 10 – 12* (Talk).
- **Hariharaputran S***, Töpel T, Oberwahrenbrock T and Hofestädt R (2007). SignAlign: Prediction and alignment of biochemical pathways using protein structural information. *In Proceedings of 5th International Conference on Pathways, Networks, and Systems, Porto Heli, Greece, 24 – 29 June* (Platform presentation – Selected abstract).

- **Hariharaputran S*** and Thoralf Töpel (2006). Structure-based alignment of linear biochemical pathways. *In Proceedings of Computational Insights into Biological Systems, Bangalore, India, 26 -28 December*. (Oral presentation - Selected abstract)
- **Hariharaputran S*** and Thoralf Töpel (2006). A Tool for Alignment and Prediction of Signaling Pathways. *In Proceedings of German Conference on Bioinformatics (GCB 2006), Tübingen, Germany, 20 – 22 September*. (Poster D8)
- **Hariharaputran S***, Chen M and Hofestädt R (2005). Alignment and classification of signaling pathways. *In Proceedings of German Conference on Bioinformatics (GCB 2005), Hamburg, Germany, 5 – 7 October*. (Poster 16)

Other publications from Sridhar Hariharaputran

- Sivapriya K, **Hariharaputran S**, Suhas VL, Chandra N, Chandrasekaran S (2007). Conformationally locked thiosugars as potent alpha-mannosidase inhibitors: Synthesis, biochemical and docking studies. *Bioorg. Med. Chem. Sep 1; 15(17):5659-65*.
- **The NMITLI – BioSuite Team**: M. Vidyasagar et al., **S. Hariharaputran** et al. (2007). BioSuite: A comprehensive bioinformatics software package. (A unique industry–academia collaboration) *Current Science Vol.92, No 1, 10 Jan; 29-38*.
- Chaitra MG, **Hariharaputran S**, Chandra NR, Shaila MS, Nayak R (2005). Defining putative T cell epitopes from PE and PPE families of proteins of Mycobacterium tuberculosis with vaccine potential. *Vaccine. Jan 26; 23(10):1265-72*.
- Prasad T, Subramanian T, **Hariharaputran S**, Chaitra HS, Chandra N (2004). Extracting hydrogen-bond signature patterns from protein structure data. *Appl. Bioinformatics ; 3(2-3):125-35*.
- Kumar P, Rao AG, **Hariharaputran S**, Chandra N, Gowda LR (2004). Molecular mechanism of dimerization of Bowman-Birk inhibitors. Pivotal role of ASP76 in the dimerization. *J Biol Chem. Jul 16; 279(29):30425-32*.

- Sivakumaran S, **Hariharaputran S**, Mishra J, Bhalla US (2003). The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics. Feb 12; 19(3):408-15.*

* - indicate presenting and corresponding author

Contents

Acknowledgements	iv
Summary	vi
Publications	vii
Contents	x
List of Tables.....	xii
List of Figures	xiii
Preface.....	xvi
Chapter 1	1
Introduction.....	1
1.1 Motivation.....	1
1.2 The Problem.....	3
1.2.1 The Goal.....	6
1.2.2 Bio-silico systems – a multi-pronged approach towards analysis and representation of complex biological data	9
1.2.3 Organization and content	11
Chapter 2	13
The Basics	13
2.1 Biological Pathways.....	16
2.1.1 Metabolic pathways.....	17
2.1.2 Signalling pathways	20
2.1.3 Protein-Protein interactions.....	23
2.2 Bio-data and integration.....	24
2.3 Molecular databases.....	26
2.4 Data warehousing.....	32
2.5 Data visualization.....	34
2.6 Pathway Biology	36
2.6.1 Pathway databases.....	37
2.6.2 Pathway alignment	38
2.6.3 Pathway reconstruction	40
2.6.4 Modelling and simulation.....	42
Chapter 3	50
Relational data integration, visualization and network navigation and exploration	50
3.1 Introduction.....	50
3.1.1 Related works.....	51
3.1.2 Bio-data warehousing approach	53
3.2 VINEdb - Visualization Integration Network Exploration	56
3.2.1 Concept	56
3.2.2 Architecture.....	57
3.2.3 Implementation.....	59
3.2.3 Application	64
3.3 Summary	69
3.4 Outlook	71
Chapter 4	72

Structure based Information and Integration for the Alignment of Biological Pathways	72
4.1 Introduction.....	73
4.1.1 Structure versus sequence	76
4.1.2 Structure versus Enzyme classification	77
4.2 Related works - alignment a common theme	77
4.3 SignAlign	81
4.3.1 Concept	83
4.3.1 Architecture.....	90
4.3.2 Implementation.....	91
4.3.3 Application.....	92
4.4 Limitations	97
4.5 Summary	97
4.6 Outlook	99
Chapter 5	100
Petri net based reconstruction and visualization of biological pathways using integrated molecular data	100
5.1 Introduction.....	101
5.2 Reconstruction, modelling and visualization of biological networks.....	105
5.3 MoVisPP	110
5.3.1 Concept	112
5.3.2 Architecture.....	112
5.3.3 Implementation.....	116
5.3.3 Application.....	119
5.4 Summary	124
5.5 Outlook	125
Chapter 6	126
Application.....	126
6.1 Construction and reconstruction of biological networks – an integrative approach.....	126
6.2 Summary	138
Chapter 7	140
Hypotheses and Discussion.....	140
Chapter 8	143
Conclusions.....	143
Chapter 9	148
Perspectives	148
Glossary	149
Bibliography.....	153
Appendix I	163
Appendix II	164
Appendix III	165
Short Vita.....	167

List of Tables

Table 1: A table showing the current statistics of KEGG databases	29
Table 2: Typical interpretation of transitions and places.....	48
Table 3: A comparative table showing the salient features of different data warehouses.....	55
Table 4: List of biological objects and corresponding public life science database integrated into VINEdb.....	59
Table 5: A comparative table showing the features of different data warehouses with VINEdb	63
Table 6: List of biological objects for search and allowed parameters in VINEdb	66
Table 7: A comparative table showing the different features of the tools developed for alignment and analysis of pathways	80
Table 8: A comparative table showing the different features of the tools developed for alignment and analysis of pathways along with SignAlign	92
Table 9: Table lists the genes and miRNA targeted disease pathways.....	137

List of Figures

Figure 1: Schematic representation explaining the kernel of the work along with concept and methods implemented which are shared by different tools.	8
Figure 2: A branched tree representation of the work flow and dissertation.	11
Figure 3: Representation of proteins as computational elements (Bray, 1995).	14
Figure 4: Global metabolism map showing the complexity of metabolic pathways. (Source: http://www.genome.jp/kegg/pathway.html).....	19
Figure 5: A view of signal transduction pathways as an integrated circuit. (Source: Hanahan and Weinberg, 2000)	21
Figure 6: A model projects the role of the MAPK cascades in cardiomyocyte hypertrophy (Muslin AJ, 2008).....	22
Figure 7: The common three-tier architecture used by most biological databases consisting of DBMS, software layer and a web interface. (Source: Stein, 2003).....	26
Figure 8: The Gene Ontology representation for the gene Bcl2 involved in cardiac and many other diseases.	28
Figure 9: An illustration of data warehouse architecture along with customized data marts (Source: www.oracle.com)	33
Figure 10: The network result from STRING database shows the protein-protein interactions for Bcl2 related to cardiovascular and other diseases from its integrated data source.	35
Figure 11: KEGG map showing the various genes including Bcl2 and the pathways associated with neurodegenerative disorders.....	37
Figure 12: The results from PathAligner shows the reconstructed pathway	42
Figure 13: Various strategies for modelling biological processes (Noble, 2003).....	43
Figure 14: Integrative framework for systems biology (Ng A et al. 2006)	45
Figure 15: Fundamental components of Petri nets.	46
Figure 16: A Petri net graph showing places, graphs and arcs (Reddy et al. 1993)	47
Figure 17: The interactive domain information system as in VINEdb. The graphical image allows navigation and exploration of the interconnected data.	57
Figure 18: The system architecture of VINEdb illustrates the flow of data upstream from the original heterogeneous data sources source layer to the web application layer in between checked by the monitor component for data up-to-dateness and then integrated into the database layer.....	58
Figure 19: A graphical representation of the integrated biomedical data in VINEdb (Fietz, 2007).....	61
Figure 20: The graphical representation of the relationship between the different data sources which are integrated into the system showing how they are connected with each other. (Fietz, 2007)	62

Figure 21: KEGG map showing the role of Bcl2 in neurodegenerative diseases with different pathways.....	65
Figure 22: Protein and gene information of Bcl2 human from Uniprot and KEGG databases	66
Figure 23: This image illustrates the results of the search showing related entity information for apoptosis regulator Bcl-2 in human and the interactive graph generated from VINEdb. The links and the graph allow further navigation and exploration.	68
Figure 24: CASP3 (Gene id 836) is connected to several pathways including MAPK (hsa04010), Apoptosis (hsa04210) and Neurodegenerative disorders (hsa01510) along with protein and disease information available within the data warehouse.	69
Figure 25: Comparative alignment method proposed by Dandekar et al., 1999	79
Figure 26: SignAlign showing the web based form where the user could interact with the system by entering the relevant PDB ids for the proteins involved in the pathways. Also the SBML file can be used for choosing the relevant PDB ids..	83
Figure 27: Schematic representation of the class (C), architecture (A) and topology (T) level in the CATH database. (Source: Orengo et al. 1997)	85
Figure 28: A detailed QSCOP information for 1t36 along with SCOP ids	86
Figure 29: PROCOGNATE information for 1t36 using SCOP based classification..	88
Figure 30: A schematic representation of the 3-tier system architecture in SignAlign	90
Figure 31: The visualization method followed by SignAlign to display the alignment result based on the SCOP classification scheme. It is supported by other information and colour legend and its significance. Also shown is the option to choose the external information.....	95
Figure 32: The related information table from different sources such as Protein Data Bank (PDB), SCOP, CATH, Gene Ontology and other external links are provided as table to the user. The table follows the similar alignment colour pattern.	97
Figure 33: The schematic representation of the top-down and bottom-up approach to systems biology. (Bruggeman and Westerhoff 2007).....	107
Figure 34: CASP3 gene related to cardiovascular disease and its interaction partners as shown in VisANT	109
Figure 35: The homepage of MoVisPP along with the guidance for the user.....	111
Figure 36: A detailed system architecture of MoVisPP describing the various layers and their action.....	113
Figure 37: The graphical representation of the data warehouse structure (Spangardt, 2007)	114
Figure 38: The figure shows the relationship between the integrated data from GO (Spangardt, 2007).....	115
Figure 39: An enzymatic reaction step generated by MoVisPP	118

Figure 40: Graphical representation of different KEGG relations as a Petri net model generated by MoVisPP	118
Figure 41: The KEGG (http://www.genome.jp/kegg/) representation of the apoptosis signal pathway explains the role of Bcl2 and p53 showing the different ways of activation of apoptosis. While the crossbar arrowheads indicate inhibition the branching arcs go to alternative as well as to concurrent successors.	120
Figure 42: A screenshot of MoVisPP showing its user friendly browser in the background. The user can interact with the system for generating their desired map by following few steps involved. Also, in the forefront is a qualitative model of apoptosis pathway generated by MoVisPP showing the network along with its supported legend.	121
Figure 43: Enlarged screenshot of the MoVisPP generated qualitative model of apoptosis pathway that is shown in the inset of Figure 40. This image is supported with integrated information from the data sources and shows the presence of the genes Bcl2 and Tp53. Every participating entity is given a unique color by the system.	123
Figure 44: A part of the Cell Illustrator image that is exported from MoVisPP and it is based on the qualitative model of apoptosis pathway.	124
Figure 45: Schematic representation of diversity of networks in tissues. (Schadt and Lum, 2006).....	127
Figure 46: An overview of signalling pathways (Source: Wikipedia- Lodish H, 2003).	128
Figure 47: Activation of MAPK and its downstream effects (Khan et al. 2004).	130
Figure 48: Human apoptosis pathway generated by MoVisPP along with the integrated information.....	133
Figure 49: Protein - protein interaction network from an integrated source for Bcl2 along with MAPK and P53.....	134
Figure 50: An enhanced protein – protein interaction network showing the associated diseases for Bcl2.	135
Figure 51: An expanded integrated network showing the protein-protein interaction networks for Human Rac1, Cdc42 and Birc6.	136

Preface

Biological data generated everyday through various high throughput experiments are huge and complex to handle. Each day passes with thousands of publications printed explaining these research works. To understand the scientific literatures is by itself a major task and further developing applications or algorithms based on these works can be yet another milestone. The biological data generated from several experiments by researchers undergo comprehensive analysis and curation everyday. This further paves way to the development of several databases and data warehouses that benefit and are again accessed by the scientific community for their day to day research. The generated and integrated molecular data can be presented to them in a user friendly manner thus allowing them to explore, analyze and navigate the data with ease instead of viewing the data as a pair of table or forcing them to browse several pages for their needed information. Pathway bioinformatics and network biology is an upcoming area of research which deals with the construction, reconstruction, modelling and simulation of biological networks / pathways, which is also an abstract representation of the integrated biochemical data. At the same time large scale modelling of the pathways is also a difficult task. So, a semi-automatic method of construction of biological pathways can be a very useful approach. And in systems biology, modelling and simulation of biological pathways has a significant place that is used to explain the underlying hypothesis using the generated biochemical data and they depend on the biochemical parameters. By this method it is possible both to predict and also to prove the behaviour of the pathways. But it is not possible to generate the quantitative parameters for the all biochemical reactions that govern the pathways. Hence a non-quantitative method can be helpful. Sometimes there is a possibility of calculating the kinetic parameters from proteins structures which is otherwise available only under normal experimental conditions. Proteins structure share more common properties or to say the proteins structure are more conserved during evolution. To decipher the evolutionary relationship among the biological pathways protein structure data can help and can bring in more details along with sequence information and enzyme based methods / approaches into the pathway phylogeny. In this work, a multi-pronged approach is proposed that deals with the methods for the analysis and representation of the multifarious biological data and explain and emphasize the concept involved. Furthermore, these methods are

combined with graphical representation that will be helpful to make things simple and comprehensible as the human eye is often used to and is more comfortable with the visual representation of the facts, which helps to understand the complexity with ease.

Chapter 1

Introduction

1.1 Motivation

To understand the very complicated networks of interacting genes, proteins and small molecules that give rise to the biological form and function is a major challenge for postgenomic biology. At the same time the large amount of protein interaction data that are available now present lot of opportunities and challenges in understanding the evolution and function. Such challenges involve assigning functional roles to interactions, separating the false positive protein-protein interactions from true positives and finally organizing these very large scale interactions into various models of cellular signalling and regulatory machinery (Sharan et al. 2005). In the past research groups were focussing on the characterization of one or few genes that were involved in a particular biological process of interest. But with the completion of many genome sequences the focus is shifted and it has opened the opportunity to formulate the questions in a different manner. It has increased the genetic programs for the study of biological systems of interest. Also, it has paved the way to address the biological questions taking almost complete sets of proteins instead of analyzing one or a few of them at a time (Boulton et al. 2001). It is not possible to address the function of living organisms satisfactorily just by looking at molecules alone, not even if the study incorporates all the molecules. The main reason, these studies would not decipher the supramolecular functional properties such as metabolic steady states, cell cycle and cell (dys) function. Furthermore, with these studies it is not possible to understand the multifactorial diseases nor would they be able to empower agricultural (green) and industrial (white) biotechnology (Bruggeman and Westerhoff, 2007). In order to understand the living cells, they must be studied as systems rather than just a set of individual molecules. It is imperative that the study of the systems is complicated as it consists of several thousands of interacting molecular species and it is necessary to have simple abstractions. Particularly it is very fruitful when the abstraction of the intracellular processes is constructed into 'networks'. The

complicated relationships between the large numbers of elements can be represented clearly with networks (Pieroni et al. 2008).

And to quote from the web pages of Institute of Systems Biology (<http://www.systemsbiology.org/>) “Systems biology is the study of an organism, viewed as an *integrated* and *interacting network* of genes, proteins and biochemical reactions which give rise to life. Instead of analyzing individual components or aspects of the organism, such as sugar metabolism or a cell nucleus, systems biologists focus on all the components and the interactions among them, all as part of one system. These interactions are ultimately responsible for an organism’s form and functions. For example, the immune system is not the result of a single mechanism or gene. Rather the interactions of numerous genes, proteins, mechanisms and the organism’s external environment produce immune responses to fight infections and diseases. Systems biology emerged as the result of the genetics "catalog" provided by the Human Genome project, and a growing understanding of how genes and their resulting proteins give rise to biological form and function. The study of systems biology has been aided by the ease with which the internet allows researchers to store and distribute massive amounts of information, plus advances in powerful new research technologies, and the infusion of scientists from other disciplines, e.g. computer scientists, mathematicians, physicists, and engineers”. The final goal of systems biology is to achieve and understand how the life forms work in an integrated manner despite of possessing unique and individualistic complexity of interactions at various levels. This is not possible to understand and imagine without an important evolutionary component; that helps to understand the changes that happen in biological systems in time and also to decipher the ways. These changes happen by selection and some of them being neutral. A change at one level of a system strongly affects on the evolution at other levels. The emergence of evolutionary systems biology is the result of the accumulation of genome-wide data pertaining to the different facets of expression of genes, their function and evolution. There are many correlations between variables that characterize the functioning of a gene, such as level of expression, knockout effect, relation between gene and protein-protein interaction networks and also between the variables that describe the evolution of gene, such as gene loss propensity and sequence evolution rate. The earlier attempts on multidimensional analysis of genomic data resulted in composite variables

describing the ‘status’ of a gene in the genomic community. However, it is not clear whether different functional variables affect the gene evolution as a single dominant factor or they affect in a synergistic manner (Koonin and Wolf, 2006).

1.2 The Problem

“Biology easily has 500 years of exciting problems to work on” said the famous computer scientist Donald E. Knuth.

What is the key to successful biological analysis? The answer for this question is that we not only understand the interactions between the basic and key components of the cells, organs and systems but also their interactions that can be affected in disease. And this valuable information resides neither in the individual proteins these genes encode nor in the genome but they are based at the level of protein interactions that is within the context of sub-cellular, cellular, tissue, organ and system structures.

Thinking about the situation does not leave any other option but to compute these interactions and also copy their nature in order to determine the logic of healthy and disease states. With the development of powerful hardware and complex algorithms computing made possible to explore the functionality in a quantitative manner, starting from the level of genes to the physiological function of whole organs and regulatory systems as a result of enormous growth of biological databases, cell models, tissues and organs.

What is the result of using powerful computing and high end technologies in biological research? Using these technologies in modern biological research obviously has provided wealth of knowledge about individual cellular components and about their functions along with enormous amount of molecular data. Moreover the information about the complex interactions between proteins, nucleic acids and small molecules which result in most cellular processes are also provided. By identifying the molecular interactions correctly, often throws light on the molecular mechanisms underlying biological processes. And it quite obvious that in order to understand or predict the overall behaviour of a complex system it is not sufficient to understand only the simple fundamental laws governing the individual building blocks. The existence of considerable variation both in the nature of elementary building blocks and their interactions requires development of novel methods that will be able to uncover cellular organization and functional principles at the systems level, is because

of the elementary forces that shape the complex and highly non-linear interactions between genes, proteins and metabolites.

Like any other large scale project where the rudimentary data is generated by different work groups across the globe by different experimental techniques and methods, for the application of this work data is generated by different research groups who are participating in the EU Project CardioWorkBench - *Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analyses* (<http://www.cardioworkbench.eu/>) is utilized. The basic data or the raw data is generated from different experiments by the project partners who are experimentalists or medical professionals from the patients in hospitals, which is further sent for micro array and other analyses to other partners in the project thus generating huge amounts of significant data. And these molecular data is again associated with different genes and proteins and their biochemical parameters and they in turn play a significant role in various diseases and also have some association with cardiovascular and other diseases. For example it has been proved by the experimental/*in vitro* methods the genes and proteins Bcl2, Rac1, BIRC6, Cdc42, Casp3 play a role in cardiovascular disease. The information about these genes and proteins are also available in different sources generated by computational/*in silico* methods e.g. KEGG, GO, OMIM etc. which again may contain information from different experimental sources as well. Furthermore, these data can be associated to hundreds of proteins, genes, enzymes, diseases, pathways and their interactions in the cell will be very complex. In order to know the relations and reactions it is not possible for biologist to perform several experiments which can be time consuming and also labour intensive. At the same time it is not guaranteed that all the experiments will be successful to generate the related data that is required for their research. This poses a problem, to know how the rudimentary data can be used and what are the methods that can be developed to decipher the relationships between the gene, protein information that are generated by the researchers and how they can be explored in order to explain their role and relationship with cardiovascular and other diseases. It is also essential to know the role of biological pathways if any, which participate in the process. Also, the problem is to find ways how this basic information can be converted to a knowledge that can be utilized for the “*Drug Design for Cardiovascular Diseases*”.

Hence there is a need for analysis and find new ways to know more about the rudimentary data and their relations with other molecular information. And this is possible by developing computational methods which can confirm as well as predict the steps and will be helpful to the experimentalists or researchers at a greater scale. By this approach they will be able to save time and money and also it is useful for them to plan the next stages of their experiments. The biochemical information are widespread and are available from different sources like scientific literature, databases etc. and data are also generated everyday in the biochemistry laboratories.

Browsing the web pages for the needed information is very time consuming. For this a long list of interactions or a protein pairs table will not suffice to show what happens inside the cell. Further the protein-protein interaction networks and similar networks generated by different applications are many times too huge and the resulting image or graph itself is sometimes too hard to understand and also confusing. At the same time the molecular data from the sources like KEGG, GO, SCOP, CATH, MINT, TRANSPATH, UNIPROT, OMIM, IntAct etc. are quite diverse and difficult to handle. Each database follows its own approach to relate and represent the data. It is required that the related information of the queried protein, gene, disease etc. be shown in a very user friendly manner in order to allow the end user to gain more knowledge.

Moreover proteins share similarities at different level when they have a common origin. Proteins join together in biological pathways in order to perform a certain task and most of proteins in different pathways across organisms are conserved, meaning they share common properties. Prediction and analysis works, to compare the biological pathways based on the different properties, can throw more light to their evolutionary relationship. And pioneering works have been done by different research groups for the analysis and comparison of the proteins at different levels i.e. sequence, enzyme and function and for structure comparison using different methods. Still these approaches could not answer some of the questions and does not utilize all the information of proteins and pathways which can be very significant for this and other research works. Having the raw data generated by the experimentalists in the project and also the know-how of the methods in which these data can be utilized and analyzed, they provoke several questions that are very significant to answer. Furthermore, it is also possible to calculate the enzymatic kinetic parameters from

protein structures which are normally required for the modelling and simulation of biological networks. The kinetic parameters are many times not available or only under experimental conditions.

In order to model and simulate biological networks, it is required to construct the network using the generated data and information. Moreover, while constructing the biological networks it is not always possible to provide the quantitative data or kinetic parameters for all the participating reactions and further elevating them for modelling and simulation using different suites. In addition, biological networks are often hard to construct manually and it is also quite cumbersome to built large networks with these properties which encapsulate integrated information from different data sources. Hence, a system that helps to construct the network in an automatic manner and also helps in the modelling and simulation of these networks without depending on the quantitative data can be very useful to the scientific community and also to the project. Furthermore, to construct networks along with other related information has a major advantage.

1.2.1 The Goal

“As exciting as the new field of genomics is, it has not yet produced a basic conceptual change in biology. The fundamental problems remain: the origin of life, cell organization, the pathways of differentiation, aging, and the molecular and cellular capabilities of the brain. What has occurred is an explosion of molecular information obtained by genomic sequences, which will soon be followed by exhaustive catalogs of protein interactions and protein function. This wealth of information can be analyzed and manipulated only with the help of computers. The rapidly expanding role of computers in biology may usher in a profound conceptual change in how we study living systems in the laboratory” (Collado-Vides and Hofestädt, 2002).

The objective of this work is to know more about the diverse and significant molecular data that are available from different resources, followed by a careful analysis from various angles and interpretation and then to present a solution to the problem using computational methods. The idea is to utilize the rudimentary biochemical data generated by the experimentalists both for the prediction and analysis i.e. the work will utilize the raw or molecular data, further mine the possible

relationships that are available among these independent data by using them in different approaches or methods which are developed by computational means. In a nutshell, the goal is to integrate the information of both *in silico* and as well as *in vitro* analyses which are very diverse in nature in order to know their significance and then to present various tools along with application case for them that can be used for the drug design for cardiovascular diseases and similar projects where these tools and methods can be used for their research.

The goal is also to answer to some of the questions that is very essential for the work like

- (a) How to integrate the multifarious biological data and what are the methods that can be adapted?
- (b) How to present the complex molecular data in an effective and comprehensible manner to the user?
- (c) Utilizing the diverse data how they can be transformed into biological networks and perform modelling and simulation of the pathways?
- (d) How to tackle the problem of non-availability of quantitative data or biochemical parameters such as the kinetic values and still construct networks effectively and project the essence of complex nature of the data?
- (e) What are the methods and information that can be helpful to know the evolutionary relationships among proteins, pathways and disease?
- (f) How much is the protein structural information useful in deciphering the homology among biological pathways especially disease related pathways?

Moreover, the objective of the work is also to analyze the multifarious biological data that are present in various sources like scientific literature, as databases and, are the results of throughput and laboratory experiments outside the project. Further, the aim is to utilize them for different works that can involve creation of data warehouses which will allow integration of the widespread data, for the methods of pathway comparison, discovering the patterns among the pathways and for the construction of large scale networks. This will be elevated and followed by modelling and simulation of biological pathways using different methods or by using the software suites.

All databases or data warehouses or tools or applications before creation have a goal or a question to answer at the end like “How” “Why” and “What”. All these questions

can be answered at the completion of the database or application bearing in mind simultaneously they have to be comprehensible to the end users who will be working with those systems or applying those concepts in their research.

The central theme or the kernel of this work will be focussing on the three most important aspects of handling the biological data.

- (a) *Integration* of the data that allows the unification of the diverse information present in different source or resources - each tool/application has its individual database as it belong to different projects that are being carried out at different times and they possess different combinations of diverse sources, which will be discussed in the forthcoming chapters.

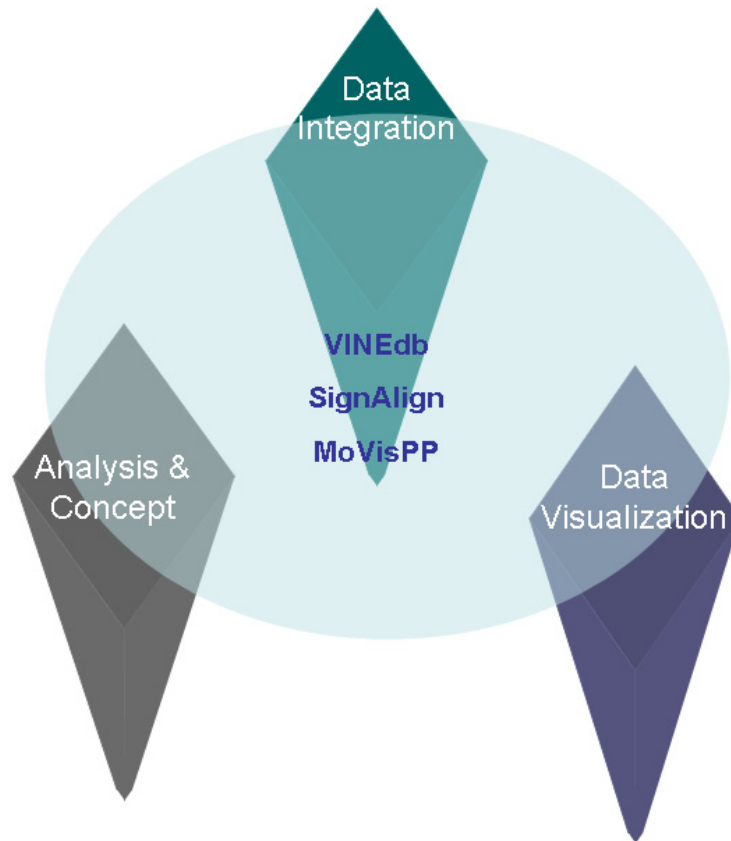


Figure 1: Schematic representation explaining the kernel of the work along with concept and methods implemented which are shared by different tools.

- (b) To emphasize on the *visualization* and representation of the integrated data in a very user comprehensible manner - all the applications or tools developed and the Unified network reconstruction approach have a common feature of projecting their results in a user friendly way using visualization techniques.

- (c) Careful *analysis and development* of new concepts to understand and to appreciate the biology and for handling the rudimentary molecular data - every application described has their own concept for the analysis of molecular data and to present their results.

Then the following tools developed will be discussed in the thesis. VINEdb, SignAlign, MoVisPP. VINEdb emphasizes on the generation of composite network generation, SignAlign find homologous pathways and to know the pattern using an alignment approach, MoVisPP helps to construct the Petri net based biological networks and reduces the time and manual intervention using semi-automatic methods. Shown in Figure 1 is the schematic representation of the integrated concepts and methods which form the base/kernel or the pillars implemented in the applications/tools and data warehouse which lie on the top. They will be discussed in detail in the forthcoming chapters.

1.2.2 Bio-silico systems – a multi-pronged approach towards analysis and representation of complex biological data

This work focuses on the development of a multi-pronged approach that is aimed to provide a solution to the problems defined in the earlier sections. Three different tools are proposed which provide a solution and can help the biologists in the analysis of the rudimentary data and also can give more information about the gene, pathways disease, etc. Furthermore, a method to expand the knowledge of the biochemical data is also proposed which can help the researchers in drug design and to know more about cardiovascular disease and can be applied to similar projects.

The whole is more than sum of its parts and the graphical representation of the facts makes it easier to understand the complex situations or biological complexity. Moreover, the use of graphics suits the human preference for visual perception. Therefore, there is a need for a system that can present a consolidated image and can project or give an overview of the background information of the biochemical data. This will help to know more about the rudimentary data which are generated by the biologists/biochemists and can help them in the next steps of their experiments. The data warehouse system developed is supported with relational visualization that

enables interactive exploration and network navigation along the complex molecular data.

An algorithm or a method which can compare the proteins and pathways using structural, sequence and other related information is proposed. In another approach the method utilizes the data for the pathway comparison and results are generated using an integrated approach. By this, it will be possible to elucidate and to answer some questions that arise out of the complexity created out of the interactions of protein and pathways and also about their relationship. A visual comparison of pathways among different organisms along with added features introduces a new method of comparing biological pathways.

A web-based tool for the automatic construction of Petri net pathways is implemented which provides a new direction towards the large scale conversion and construction and reconstruction of biological pathways. The tool will be able to construct pathways using the rudimentary data using the diverse information that are available from different sources. The generated pathways can be further exported and modelled using various Petri net based simulators. These networks are not dependent on the quantitative or biochemical parameters that are required for their construction by other methods.

Furthermore by combining the results generated from these tools a method to construct / reconstruct the networks is proposed. This network is generated using an integrative approach. The reconstructed network combines various aspects of unifying the rudimentary data that is of medical significance with the protein-protein interaction, ontology, disease and further more information. This is a significant approach in network biology that uses both *in silico* and *in vitro* analyses and results which can help in the drug design for cardiovascular diseases and can be applied to other similar projects.

This work is a multi-pronged approach for the analysis and presentation of the multifarious biological data. And the flow of work and different chapters of the thesis is projected as a branched tree in **Figure 2**. *Chapter 1* begins with an introduction of the topic, further proceeding and focusing on the problem and will discuss the goal of the work and *Chapter 2* deals with the basics. *Chapter 3* presents a data warehouse which is based on the concept of relational visualization, integration, network

navigation and exploration which discusses about the composite network visualization with embedded information of proteins, pathways, diseases etc. *Chapter 4* will deal on the pathways discussed in previous chapter and will discuss about a comparison

1.2.3 Organization and content

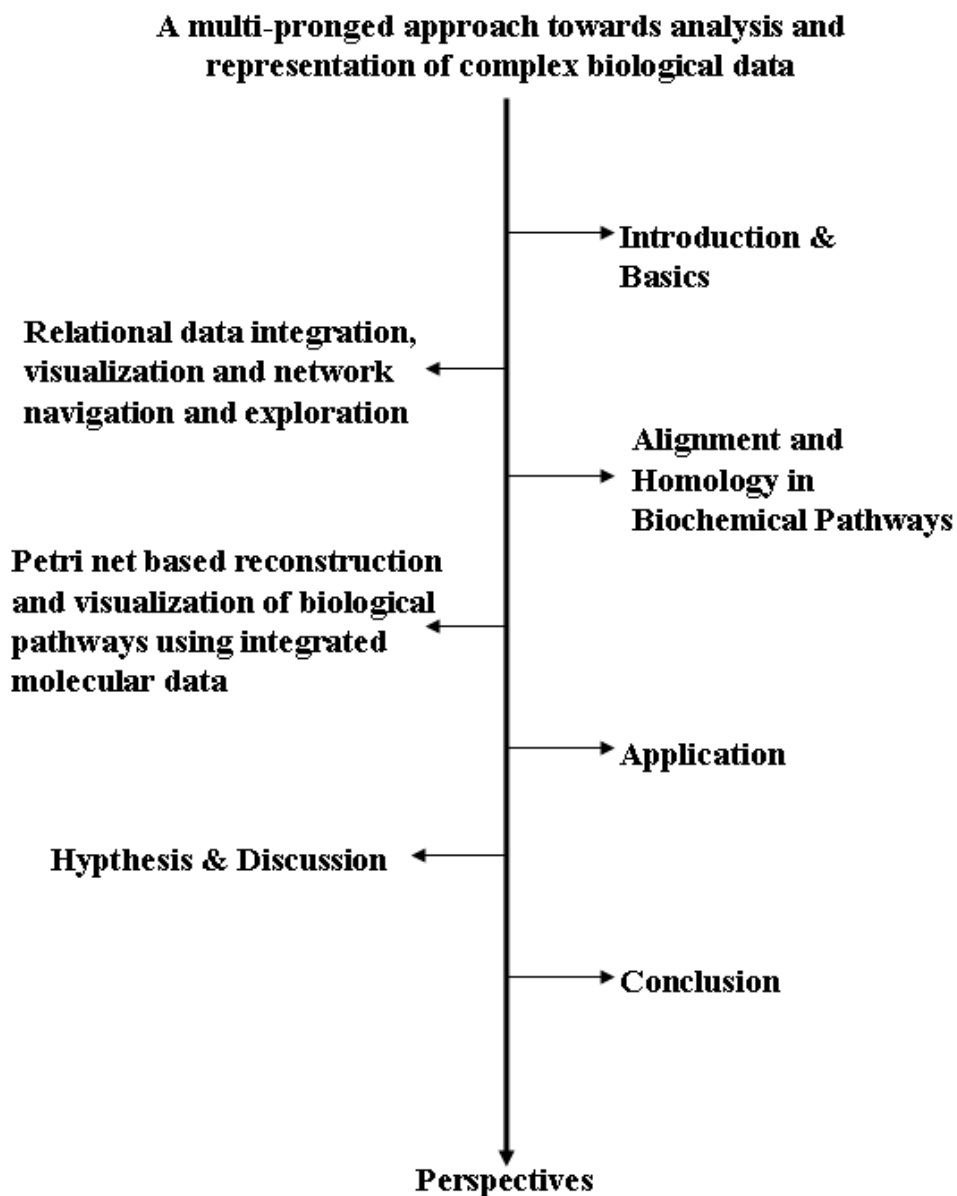


Figure 2: A branched tree representation of the work flow and dissertation.

approach and alignment of biological pathways using protein structural information and classification. This is followed by a Petri net based approach to construct large scale biological network using semi-automatic methods which is introduced in *Chapter 5* which is built using integrated molecular data. *Chapter 6* deals with the application of the works further proposing a hypothesis and network that will be

useful for the analysis of data produced by CardioWorkBench and similar projects. *Chapter 7* summarizes the underlying hypothesis for each application / chapter discussed before. Finally, *Chapters 8 and 9* will deal with the conclusions and perspectives about the works. They are followed by glossary, *bibliography and appendix*.

Chapter 2

The Basics

Living cells are subjected continuously to various stimuli from their environment, which require suitable responses. For maintaining the stability of their internal condition or to perform cellular functions, these cells are powered with biochemical machinery of which proteins are the key molecular entities. The biochemical machinery is highly characterized at the molecular level and efforts to understand their large scale organization is made by the systems biology community (Alberto de la Fuente et al. 2008). It is known that growth and differentiation and functions of the human cell are controlled by the actions of thousands of protein. Seldom does the human protein work alone. Instead they join together with other proteins to form complexes in order to exert their function in concert. During different biological processes such as gene transcription, DNA replication and repair and during other processes functionally related protein complexes interact together forming specific cellular machineries. The situation is further complicated, making the network of protein interactions sustaining cell growth and differentiation a dynamic assembly along with a very complex organisation, as in polypeptide which could assemble into more than one protein complex (Coulombe et al. 2008).

Proteins are large organic compounds. The primary structure of proteins is represented by the amino acids along with the bonds. These amino acid residues occur in a certain order which is called the amino acid or peptide sequence and the residues are bound by a peptide bond. This primary structure undergoes modifications leading to the secondary, tertiary structures. Normally, the gene defines the sequence of amino acids encoded in the genetic code. To achieve a certain goal or function, proteins can work together. Furthermore many proteins are enzymes and they catalyze the biochemical reactions that are very important to metabolism. Many of the biological

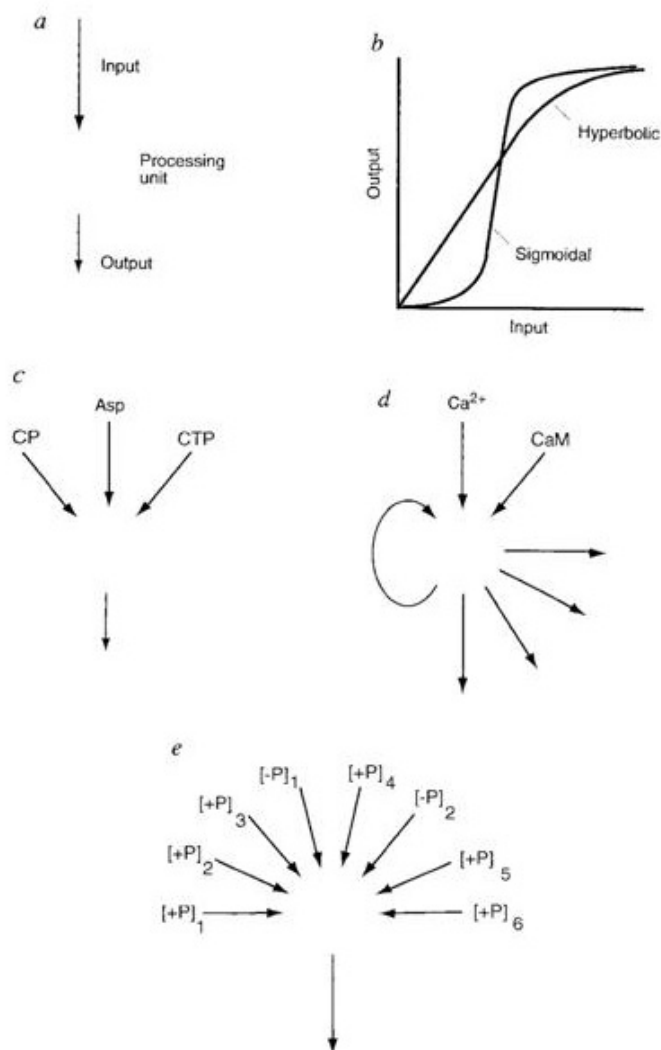


Figure 3: Representation of proteins as computational elements (Bray, 1995).

activities are controlled and mediated by proteins. They function as single units or mostly associate with other partner molecules or become part of a larger assembly. And understanding them alone and how they act in unison with other proteins is necessary. The cells not being static are prone to many external and internal signals that can alter the structure, shape and other properties in the proteins involved. The cellular functions vary within (intra) and outside the species (inter species) (Golemis E.2002). Further it appears that most proteins in living cells have the transfer and processing of information as their primary function, more than just chemical transformation of metabolic intermediates or cellular structure building. The proteins are functionally connected through allosteric or other mechanisms into multi-tasked biochemical “circuits” and perform a variety of simple computational tasks (**Figure 3**) consisting of amplification, integration and also storage of information (Bray, 1995).

The chemical and structural properties of proteins allow them to communicate and act according to the changes that happen in their ambience. These changes or the perturbations stimulate a myriad of components upstream or downstream. The change can be structural, chemical or physical. This in turn produces a biological answer treating the stimulus as a question. The answer once again can be anything, from cell division to cell death. Here it poses yet another question, to know how it is possible for a complex biological system to function effectively despite all perturbations and with robustness.

A gene is not an island. Genes work together and also take part in many biological processes that determine the cell's behaviour and phenotype, even in a single cell. Cellular networks such as signalling networks, gene regulatory networks and metabolic networks are due to the interdependent interactions of genes and proteins in cells. Cancer complexity is caused due to the genomic alterations and for understanding the molecular mechanisms that cause oncogenesis it is a major hurdle (Wang et al. 2007). An important question that arises once a genome is sequenced is to know the presence or absence of biological pathways. The analysis of biological pathways within a genome is a complicated task as more number of biological entities is involved in them. Moreover biological pathways involved are not identical across different organisms. Thus, computational pathway analysis and identification involve different variety and number of tools and databases and is practised by the comparison of pathways in many other organisms. As the requirements of computational methods are beyond the capacity of biologists there is always a need for information systems helping them for the reconstruction, annotation and analysis of biological pathways (Choi and Kim, 2008).

In order to execute the essential functions in the cells such as replication, transcription and protein transport the proteins combine into macromolecular complexes. A functional module of a protein is defined a set of proteins which are related by one or more genetic or cellular interactions e.g. coexpression, coregulation being a member of protein complex, signalling or metabolic pathway or a cellular aggregate (chaperone, ribosome, protein transport facilitator, etc.). A significant property of the module is that the function of module is separable from other modules and that the relationship is more among inter modules than intra modules (Li et al. 2007). All biological components from individual genes to entire organs work in concert in the

human body in order to promote normal development and sustain health. The extraordinary feat of biological teamwork is achieved because of the battery of intricate and highly intertwined pathways that facilitate among genes, molecules and cells. Biological pathways, while some of these have already been discovered, yet there are lots of hidden information and more to be explored. Moreover to understand how these pathways are united in human and several other complex organisms, research is needed. And to decipher how the disturbances in these pathways may lead to disease and to restore the disturbed pathways to their normalcy, research as well is needed.

2.1 Biological Pathways

The word “pathway” means very different things for different researchers. A pathway can refer to a metabolic pathway that involves a sequence of enzyme catalyzed reactions or a signalling pathway comprising of several protein-phosphorylation reactions and gene regulation events. In the past 5 to 7 years pathway bioinformatics has gained more importance and is an active area of research (Green and Karp, 2006). The actions of genes, proteins and metabolites in concert are often conceptualized as pathway diagrams. Biological pathways represent a concept of biological research as they are useful abstraction of key ideas of biology and also help to organize the researcher to answer the research questions using software tools dealing with pathway diagrams (van Iersel et al. 2008). Networks of complex reactions at the molecular level in living cells are represented as biological pathways. They show how biological molecules interact to accomplish a certain biological function and their response to the external or environmental stimuli. Further pathways are derived from data analysis and scientific experimentation and help to capture the recent knowledge (Sariya et al. 2005).

A pathway has a special biological meaning and it is a part of a biochemical network. It corresponds to a subgraph of the whole graph according to the graph theory. A pathway may contain branches and joints and it is not necessarily linear. Glycolytic or the pentose phosphate pathways are well known pathway examples. Though not necessarily related to the whole biological functional unit consequently a subpathway is again a part of a pathway (Grafahrend-Belau et al. 2008). According to Pathguide, the pathway resource list, there are **310** (as of December 2009) resources containing

information about protein-protein interactions, metabolic pathways, signalling pathways, gene networks and others (Bader et al. 2006).

Historically cell biology has focussed on the individual interactions between genes, proteins, RNAs and metabolites and to know how the phenotype is affected by each interaction. Known as *pathways*, the chains of casual interactions mediate the signals that travel around and between each cells that essentially controls the phenotypic behaviour. Very less is known about the structure of the pathways which comprises large numbers of interactions and it is very expensive and time consuming to conduct experiments for studying the constituent individual components. Biological pathways constitute one of the following: metabolic pathways, molecular interactions, gene regulatory networks and signalling pathways. Historically they bounded different areas of expertise and the boundaries which exist are entirely artificial between these seemingly distinct categories. It is defined that a pathway must have an input (the starting state) and an output (end state) and the main aim of pathway biology is to determine the relationship between the two states (Watterson et al. 2008).

2.1.1 Metabolic pathways

It is not sufficient only to understand the chemical compositions and three dimensional structures of biological molecules in order to understand how they are assembled into organisms or to understand how they function to sustain life. It is also essential to examine the reactions in which the biological molecules are built and how they are broken down. Moreover we should also know how the free energy is consumed in building these cellular materials and how the energy is generated from the organic resources. The organisms acquire and use free energy for carrying their functions through the process of “Metabolism”. The metabolism is again divided into two categories. (a) *Catabolism* or the degradation process which breaks down the cell constituents and nutrients to salvage their components and / to generate energy.

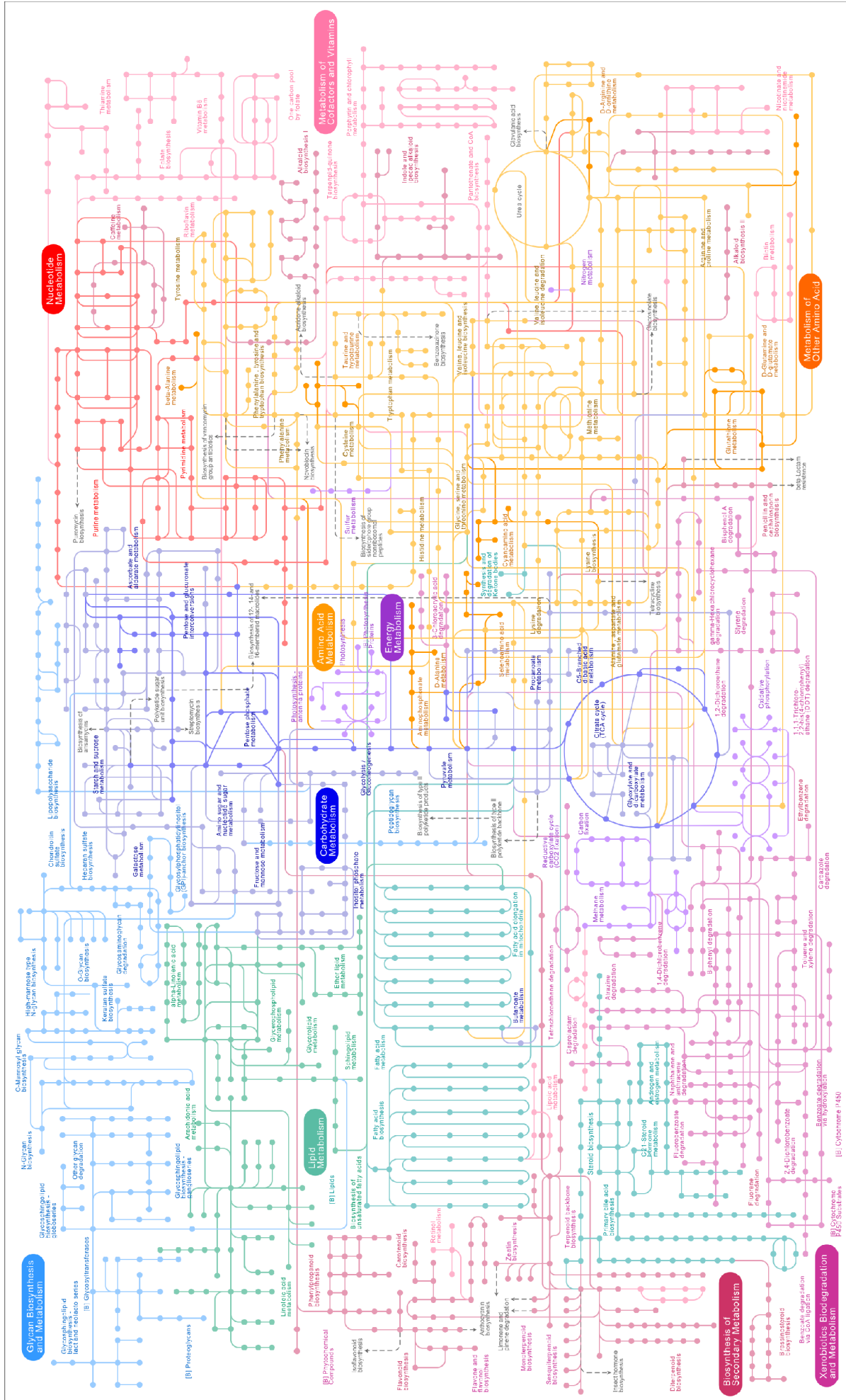


Figure 4: Global metabolism map showing the complexity of metabolic pathways.

(Source: <http://www.genome.jp/kegg/pathway.html>)

(b) *Anabolism* or the biosynthesis process in which simpler components synthesises biomolecules. Though an array of chemical reactions occurs in any living cell in every organism the principles that govern metabolism are the same. This is due to the constraints of the laws of thermodynamics and due to their common evolutionary origin. With differences primarily to different sources of free energy that support them, many of the specific reactions of metabolism are common to all organisms (Voet et al. 1999). It has been a long standing challenge for biologists to understand the complex metabolic processes of organisms. Catalyzed by special purpose proteins these processes consist of map of chemical reactions. Complexity of the metabolic system further motivated the creation of metabolic pathway, an abstraction, in order to provide a simpler view of the complex network. To segment the map of metabolic processes into logical sections of related reaction, the metabolic pathways pave the way. Evolution generates some organism-specific variations although two species may have similar metabolic pathways (Pireddu et al. 2005).

“Metabolic pathways are series of connected enzymatic reactions that produce specific products”. Metabolites constitute their reactants, intermediates and products. There are several thousand metabolic reactions and each of them is catalyzed by a unique enzyme. With the identity of the organism, the cell type, its nutritional status and its developmental stage the types of enzymes and metabolites vary in a given cell. Delineating a pathway from a network that comprises of several thousands reactions is somewhat arbitrary and further driven by tradition as much as chemical logic. Many metabolic pathways are branched and highly interconnected (Voet et al. 1999). Moreover it is a series of individual biochemical reactions, connected through substrate and product metabolites, that produces a set of metabolites from a set of precursor metabolites and cofactors. Metabolic pathways concentrate on the conversion of small molecules, and the enzymes involved in the conversion of these small molecules, making them different from signalling and protein-protein interactions pathways and maps. Different metabolic pathways are allowed to operate in different locations due to the compartmentation of the eukaryotic cytoplasm. The global metabolism map of KEGG (**Figure 4**) and the famous wall chart of Roche

(http://www.expasy.ch/cgi-bin/show_thumbnails.pl) provide an overview of the complexity of metabolic pathways.

Among the pathways, metabolic pathways are studied intensively by the experts of biology, computational biology / bioinformatics and in the areas of medicine. But at the same time it is of no wonder that these pathways are also not understood properly and considered just as a combination of biochemical reactions which enable to sustain life. Thinking from a global perspective, a metabolic pathway is a just a subset of complex reactions which tells us how the given reactant is converted to a desired product and explaining the biochemical processes involved. And thinking from a graph theory perspective, there is a beginning and end point and they often can be depicted in terms as vertices, edges etc.

2.1.2 Signalling pathways

Being stimulated, the cell responds by a biological answer (**Figure 5**). In simple words it is the relationship between the response to stimuli or vice versa. But to describe in detail, it is the process or myriad of changes that occurs within the cell to provide a response when it is stimulated by an external stimuli that can be physical or chemical. Based on the stimuli the receptor in the cell is activated which further passes on the information downstream through a cascade of proteins that finally ends up with the regulation of transcription or other cellular processes such as growth, differentiation and apoptosis. For achieving this task one has to bear in mind that it does not happen in a trivial fashion but involves lot of decision making or mind boggling changes such as the transfer of the information or stimuli to the receptors that are present in the outer membrane of the cell. Further, these receptors transfer the message to the proteins that are present in the downstream, which through the process of dimerization and post translation modification passes on the information and creates a response in the form of biological answer and affect the whole cell. For these to happen there are several molecular changes that do take place such as the changes in the concentration of the enzymes present and the secondary messengers such as Calcium which either inhibits or exemplify the response. This can be called as negative or positive feedback depending on the situation. Sometimes these types of changes do produce some oscillations in the response for longer time. This type of exchange of information across the boundaries of the proteins, enzymes involved is called signal transduction. Since, this type of information happen in a methodical way

or network form, this is referred to as signal transduction pathway or signal transduction network.

The key decisions of a cell such as to differentiate, attach, move, survive or die comprises the course of development of an organism. These decisions are coordinated both by external factors and internal cues making possible the process of signal

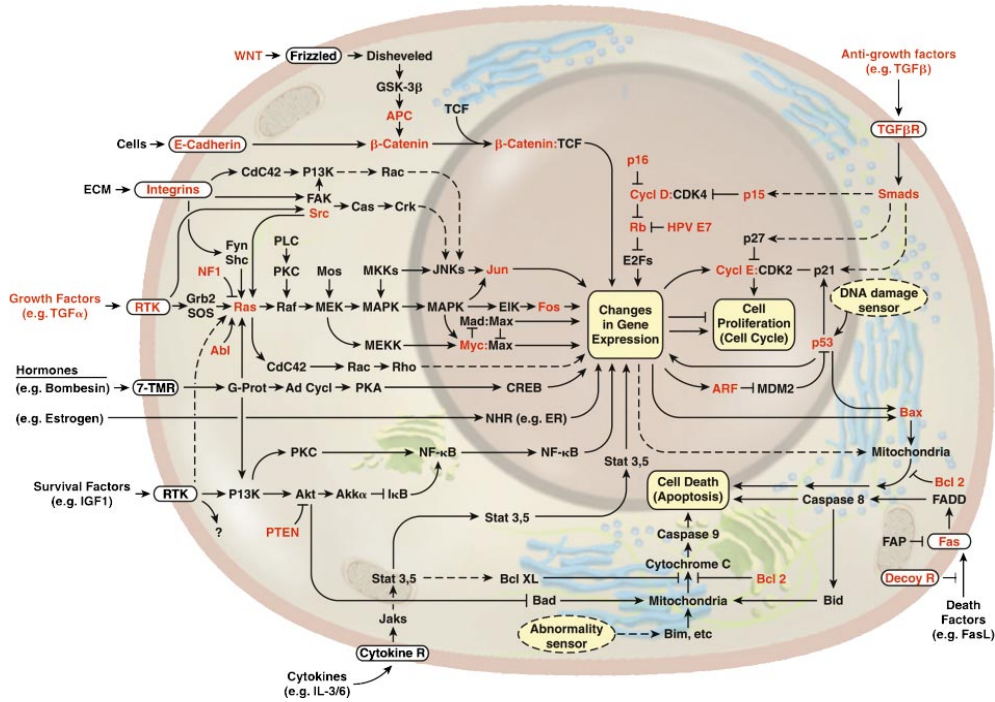


Figure 5: A view of signal transduction pathways as an integrated circuit. (Source: Hanahan and Weinberg, 2000)

transduction. Unlike the metabolic pathways which concentrate on the conversions of small molecules and enzymes responsible for these conversions, the signalling maps or pathways concentrate on the physical interactions of the protein involved without clear chemical conversions. The cell signalling is where a stimulus (light, chemical or physical) from the external environment propagates downstream by causing changes in the chemical and structural level of the receptors and the effectors that are involved. Evolutionary relationship provides a key to the understanding of signalling proteins that are conserved across different eukaryotes such as yeast, worms or flies to decipher the similar pattern or design in humans. The signalling, though complex, are just a series of simple coordination or interactions of individual protein with one another (Golemis, 2002). Signalling pathways can be thought and viewed as a module where several inputs produce multiple outcomes taking their effects through intertwined networks (Qi et al. 06).

In biological and medical sciences signal transduction pathways are of a special interest. Disturbances in signalling pathways lead to several diseases. For example, important regulators of intracellular signal transduction pathways are protein tyrosine kinases and in metazoans they also play an important role for mediating development and in multicellular communication. Perturbations of the normal autoinhibitory

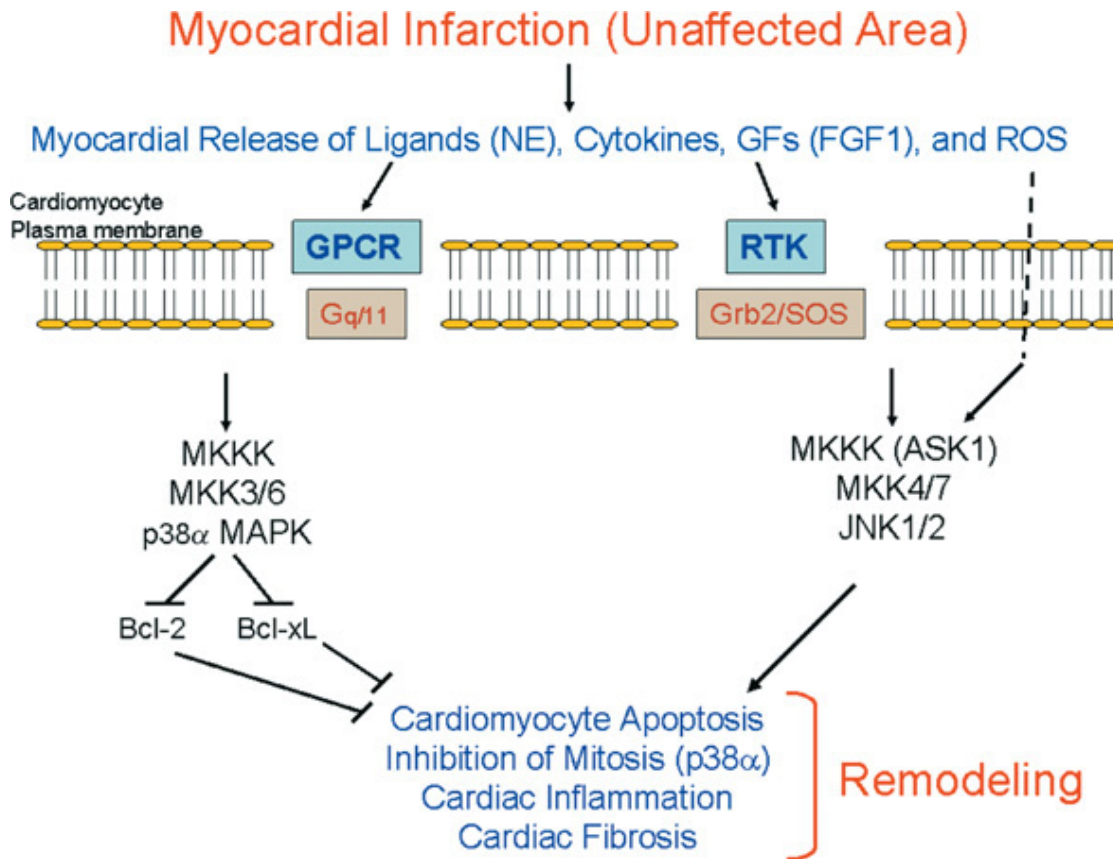


Figure 6: A model projects the role of the MAPK cascades in cardiomyocyte hypertrophy (Muslin AJ, 2008).

constraints on kinase activity can result in oncogenic PTK signaling at times, though their activity is tightly controlled and regulated. Human G-protein coupled receptors (GPCRs) which mediate response to odour, taste, light, hormones and neurotransmitters are another example. Antihistamines, β -adrenergic receptor blockers and serotonin-reuptake inhibitors are widely prescribed drugs which bind to specific GPCRs. From yeast to human there are several signaling modules occurring in eukaryotic organisms e.g. Mitogen Activated Protein Kinase (MAPK cascade) and the G-protein. Some special properties are exhibited by signaling pathways. In signaling pathways, there is a signal flow instead of substance flow being performed by different protein forms, phosphorylated and dephosphorylated and this is in

contrast to metabolic networks (Sackmann et al. 2006). It is probable that the intracellular MAPK (Mitogen Activated Protein Kinase) signaling cascades play an essential role in the pathogenesis (**Figure 6**) of cardiac and vascular disease. Even though various basic science research works has defined and deciphered the role of MAPK pathway organization and activation, it is yet to define or elucidate the role of individual signaling proteins in the pathogenesis of various cardiovascular diseases. This can lead to discovery of attractive targets of pharmacological therapy (Muslin AJ, 2008). The detrimental effects leading to the uncontrolled cell proliferation similar to what happens in cancer, the attack of other cells in the body similar to auto-immune diseases are the significant results of the malfunction of cellular signaling processes. It is not possible to understand the dynamics of cell signalling mechanisms till now which are profoundly complex. Hence in the scientific study of the biological processes which control and regulate cellular function, modeling of cell signal transduction is a significant tool (Calder et al. 2006).

2.1.3 Protein-Protein interactions

Seldom, proteins act in isolation rather they act in combination. The different levels of complexity increases when there is combinatorial interactions between them not just by the presence of number of proteins or genes (Aytuna et al.2005). Proteins, being complicated and among the most important molecules in the cell, have received the attention of biochemists for several years. Protein-Protein interaction studies are carried out for the past 30-50 years also leading to proteomic studies and huge amount of data. The generated data called on the assistance of computer science to decipher the data not possible by humans. Similar to the signal maps the protein-protein interaction maps also concentrate more on the physical interactions than chemical conversion involved in the process of protein and enzymes and their products (Golemis, 2002).

The proteins, DNA and other molecules form the components of cellular networks and they act in concert to carry out the biological processes and further interact with one another to form modules in the network. One of the major challenges in biological research is to understand how the phenotypes and behaviour of cells are controlled. The focus though has been given to the characterization of individual genes / proteins or interactions during cellular events; they cannot be attributed to isolated components (Qi et al. 2006). Researchers were able to investigate evolutionary processes in a very

detailed manner because of the various works that lead to new insights into the molecular architecture that underlie the complex cellular phenotypes. For an increasing number of species, protein interaction data are available and there are several studies that investigate the evolutionary aspects of the network interaction (Stumpf et al. 2007). A living cell is a dynamic system, where the interactions and gene activities exhibit temporal profiles and spatial compartmentalization. Cellular networks to enable their functionality contain characteristic topological patterns. Some components in the basic building blocks of the networks are found to be over represented significantly. These recurring units are defined as “network motifs”.

2.2 Bio-data and integration

Data about proteins, genes and nucleic acids and their information are integrated to create very large scale databases, pathways and networks. Further, metabolic, signalling and regulatory pathways are important for the analysis of cellular behaviour and evolution. In-silico approach can help to analyze and understand how proteins and pathways function and interact with each other. Understanding the principles of biological systems and processes is the major goal of biology, computational biology / bioinformatics. And the exploration of the enormous amount of data on various molecules and their interactions with their counterparts from day-to-day experiments followed by careful analysis and annotation by experts has paved the way to create large databases. Both high and low-throughput and post-genome experiments have further contributed to the growth of these databases with huge information. This exponential growth of biological information has persuaded the interest across different research faculty to compare these diverse data i.e. sequence, structure, expression arrays and pathways from different perspectives. Bringing these large amounts of information together is a major task, which is possible through database integration methods, and it is important.

And in order to understand and follow the scientific output produced regarding a single disease such as cancer, a few dozens of paper per day have to be read along with the scanning of more than hundred different journals by a scientist. The data that underlie show a great deal of complexity and dynamics and are produced by numerous application areas that are heterogeneous in nature. The integration of this heterogeneous data is gaining importance each day. At present different biological

data types such as sequences, protein structures and families and other experimental data along with ontologies and expression data are stored in unique databases. Existing databases can be very specialized and they also store information using certain formats. Most of the databases have not exactly the same but overlapping information that introduces another hurdle to combine the information. For gaining insights into the biological complexity and their dynamics, the information that is stored in different repositories needs to be connected and combined in an efficient manner. So, data integration has become a major topic in the past years to address these issues (Pavlopoulos et al. 2008). Several thousand scientific papers are published every year to correct old information and also to add new details about millions of unique biological entities. Among several others one of the major challenges faced by the research community is how to access, analyze and visualize heterogeneous data in ways that can lead to novel insights into biological processes or to the formulation of a hypothesis and can be experimentally tested (Blake and Bult, 2006).

From a theoretical point of view data integration is the problem of combining the data available or placed at different sources thereby providing the end user with an integrated or unified view of these data (Lenzerini 2002). Data integration which gained the appreciation by integrating petabytes of accumulated data and their technical management has grown from that level and it shifted its focus to the informational value of the data. It now aims to answer complex questions such as how a genetic background or exposure to some of the external stimuli or environmental factors can influence the development of certain diseases. The projects differ from a simple integration of similar data sets to interdisciplinary integration (Brazhnik and Jones, 2007). The omics (proteomics, metabolomics, kinomics) data that are generated by various techniques are stored in different databases and can be accessed via internet for free or are available commercially. By combining different data sets several important questions can be answered. These are performed using database integration techniques and by creating databases/ data warehouses.

Biological databases in the past two decades has grown up from a small scale industry having interest for fewer disciplines turning into important resources that are used daily by biologists around the world. There are profound examples and include some of the diverse databases such as PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>) that

allows the faster search of biological literatures maintained by the National Center for Biotechnology Information (NCBI), the Gene Ontology (GO) (<http://www.geneontology.org/>) database of gene function, process and location terms. It is sometimes hard to imagine everyday research in biology without accessing these databases. The databases differ in their functions and their information yet they

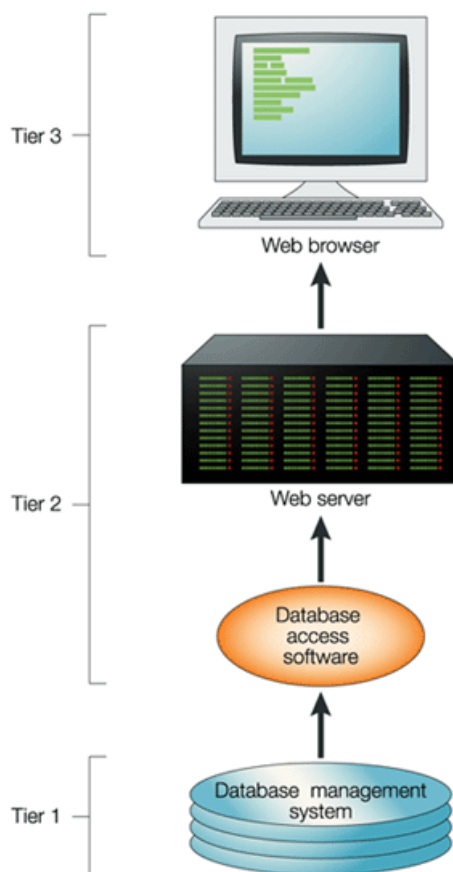


Figure 7: The common three-tier architecture used by most biological databases consisting of DBMS, software layer and a web interface. (Source: Stein, 2003)

share similar basic architecture (**Figure 7**). The bottom layer is a database management system (DBMS) that manages a collection of information. On the top of this layer is the middle tier or the software that mediates between the DBMS with the top layer i.e. the web browser. The software turns data requests into database queries, and further transform the query responses into database queries and also to transform the query requests into hyper text mark-up languages. The web browser transmits the user's requests for the data to the underlying database and renders the responses as web pages (Stein, 2003).

2.3 Molecular databases

There are several hundreds of databases dealing with variety of information and comprises both manually and automatically curated data of the biological molecules. Each of these data deals with significant information about the protein sequence or structural information, disease, pathway, ontology and protein interaction etc. Some of the information can be redundant or unique according to the database. Everyday the list is increasing adding new databases and information systems by different research groups and organisations for various purposes. The journal, *Nucleic Acid Research* (NAR) typically summarizes and categorizes the papers submitted for their yearly special database issues as follows

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases

The papers submitted are available online for free to the community and more details can be obtained from the web page <http://www3.oup.co.uk/nar/database/c/>

Even though hundreds of databases are available it is not possible to elaborate them all. Only few databases which are connected to the work and have been integrated in various applications that are developed are discussed in this chapter and in the forthcoming chapters. Each of these applications utilizes the information from these databases and presents their information in a fortified way. Also it is necessary to mention at this point that every tool has its own integrated database since they are created at different times using different concepts and methods. Each of these systems does have similar data or information from the same data source still they are part of

an individual integrated database / data warehouse. And they only satisfy the needs of the individual application or tool to which they are associated.

The Gene Ontology

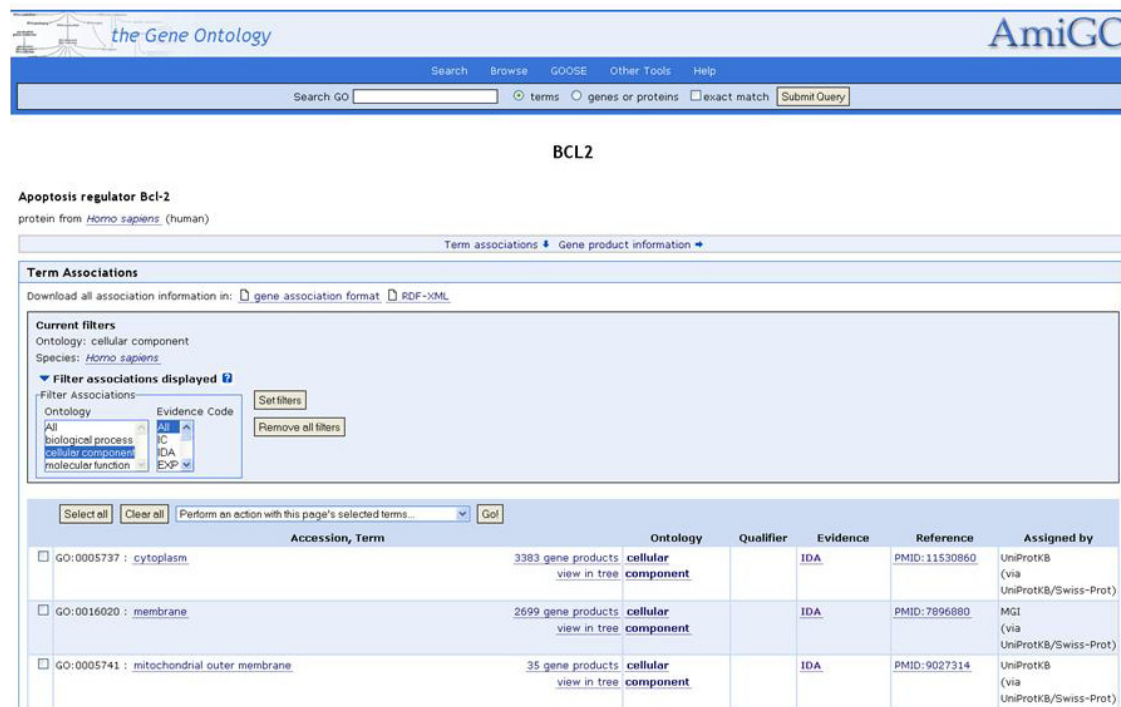


Figure 8: The Gene Ontology representation for the gene Bcl2 involved in cardiac and many other diseases.

It has been made clear by genomic sequencing that all eukaryotes share a large fraction of the genes specifying the core biological functions. The knowledge of gene and protein roles in cells is accumulating everyday and they also change. The Gene Ontology project is a combined effort for dealing with consistent descriptions of the gene products in different databases. The consortium maintains a relational vocabulary in a hierarchical (**Figure 8**) manner and describes the ontology of the gene products on three well organized levels, Biological processes “refers to a biological objective to which the gene or gene product contributes”, Cellular components “refers to the place in the cell where a gene product is active” Molecular functions “is defined as the biochemical activity of a gene product” in a species in an independent manner. The keyword databases of many gene and protein are linked to each node in the GO ontologies along with other kind of information. The aim of the project is to create a dynamic and controlled vocabulary that can be applied to all the

eukaryotes even when the roles of the gene and protein is changing and still accumulating (The Gene Ontology Consortium, 2000).

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a knowledge database. It stores the data of gene function performing systematic analysis. Further it links the genomic information with high order functional information. It integrates genomic, chemical and systemic functional information. There are several databases in KEGG. The GENES database stores the gene information, PATHWAY database contain higher order functional information along with the graphical representations of the cellular processes. The LIGAND database consists of chemical compounds, enzyme molecules and enzyme reaction information. Along with these are KEGG BRITE that stores the ontology information, KEGG drugs contains the approved drug information, KEGG disease stores the disease information linking to the genes, pathway, drugs and other relevant information. KEGG pathway is supported with supplement global maps (Kanehisa et al. 2000, 2008). **Table 1** shows the current statistics (October 2009) of KEGG database and the information about the individual databases available from <http://www.genome.jp/kegg/docs/statistics.html>

Table 1: A table showing the current statistics of KEGG databases

KEGG PATHWAY	Pathway maps, reference (total)	336 (95,697)
KEGG BRITE	Functional hierarchies, reference (total)	83 (24,814)
KEGG MODULE	Pathway modules	692
KEGG DISEASE	Human diseases	104
KEGG DRUG	Drugs	9,082
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	12,551
KEGG GENOME	KEGG Organisms	1,173
KEGG GENES	Genes in high-quality genomes (103 eukaryotes +907	4,854,441

	bacteria + 67 archaea)	
KEGG SSDB	Best hit relations within GENES Bi-directional best hit relations within GENES	30,028,627,550
KEGG DGENES	Genes in draft genomes (11 eukaryotes)	515,586,434
KEGG EGENES	Genes as EST contigs (85 eukaryotes)	151,662
KEGG COMPOUND	Metabolites and other small molecules	3,350,468
KEGG GLYCAN	Glycans	15,949
KEGG REACTION	Biochemical reactions	10,969
KEGG RPAIR	Reactant pair chemical transformations	8,021
KEGG ENZYME	Enzyme nomenclature	11,761

The Universal Protein Resource (UniProt)

UniProt formed by the unification of the activities and information contained in Swiss-Prot, TrEMBL and PIR aims to provide the scientific community with a comprehensible, high-quality, freely accessible and a single centralized, authoritative resource for protein sequences and functional information that is highly essential for modern biological research. The Consortium consists of groups from European Bioinformatics Institute, the Protein Information Resource and the Swiss Institute of Bioinformatics. Manual curation of protein sequences assisted by computational analysis, sequence archiving, provision of additional value-added information through cross-references to other databases and a user friendly website are the core activities of the consortium. Each optimized for different uses the UniProt comprises of four major components, the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. The key achievement of the Consortium in the recent past is the completion of first draft of the complete human proteome in UniProtKB / Swiss-Prot. Every three weeks the UniProt is updated and distributed and can be accessed online at the

following website <http://www.uniprot.org/> (Apweiler et al. 2004, The UniProt Consortium, 2008).

Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man (<http://www.ncbi.nlm.nih.gov/omim/>) is a database of consistently updated catalog of human genes and genetic disorders. The database mainly focuses on inherited or heritable genetic diseases. OMIM is a compiled authoritative, comprehensive and timely knowledge database of human genes and genetic disorders for supporting human genetics research and education and also for the practice of clinical genetics. OMIM is started by Dr. Victor A. McKusick as the definitive reference Mendelian Inheritance in Man is now being electronically distributed by the National Center for Biotechnology Information and further integrated with Entrez databases suite. With inputs from scientists and physicians around the world OMIM is written and edited at John Hopkins University. Each entry has numerous links to other genetic databases such as DNA and protein sequence, general and locus specific mutation, Pubmed references, Map Viewer, GeneTests, HUGO nomenclature, patient support and many others along with full text summary of a genetically determined phenotype and / or gene. It is a straightforward and easy portal to the burgeoning information in human genetics (Boyadjiev 2000, Hamosh et al. 2005).

MINT: a Molecular INTERaction database

The aim of the Molecular INTERaction database is to store in a structured format the known information about molecular interactions (MIs) thereby extracting experimental details from various published works in peer-reviewed journals. MINT is a database that is designed to store the significant data about the functional interactions between proteins. It was also conceived to store the various types of functional interactions that include enzymatic modifications of one of the partners, beyond cataloguing binary complexes. In its entry MINT aims at being exhaustive in the interaction description and also about the information about kinetic and binding constants and about those domains participating in the interactions, whenever available. MINT comprises of entries extracted by expert curators from the scientific literatures and they are further assisted by software “MINT Assistant” that presents to the curator in a user friendly format the targeted abstracts, containing the interaction

information. Further the “MINT Viewer” helps to easily extract the data and view it graphically. HomoMINT, is the database integrated into MINT recently. The database comprises of interactions between human proteins which are inferred from various experiments along with ortholog proteins in model organisms. MINT is available at <http://mint.bio.uniroma2.it/mint/Welcome.do> (Zanzoni et al. 2002, Chatr-aryamontri et al. 2007).

2.4 Data warehousing

"A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions." W H Inmon

A data warehouse houses consistent, clean, a standardized and integrated form of data sourced from different operational systems. Even though data warehouse focuses on data storage it is also meant to retrieve and analyze the electronic data and also to focus on the ETL process i.e. Extraction, Transformation and Loading of data and manage the data dictionary which is also considered being an essential component of data warehousing system. By having a universal data model it solves the problem of dissimilarity in the data model of the source databases that may adhere to its own model. By this approach it identifies and resolves the inconsistencies and simplifies the reporting and analysis prior to uploading the data. A data warehouse can be referred as a storehouse of integrated data from different sources which are being processed for storage in a multidimensional model. This is in contrast to multiple or multidatabases that often allow the access to heterogeneous and disjoint databases. Following are the unique features of data warehouses (Elmasri and Navathe, 2000)

- generic dimensionality
- multidimensional conceptual view
- unrestricted cross-dimensional operations
- unlimited dimensions and aggregation levels
- dynamic sparse matrix handling
- accessibility
- transparency
- client-server architecture

- multi-user support
- intuitive data manipulation
- consistent reporting performance
- flexible reporting

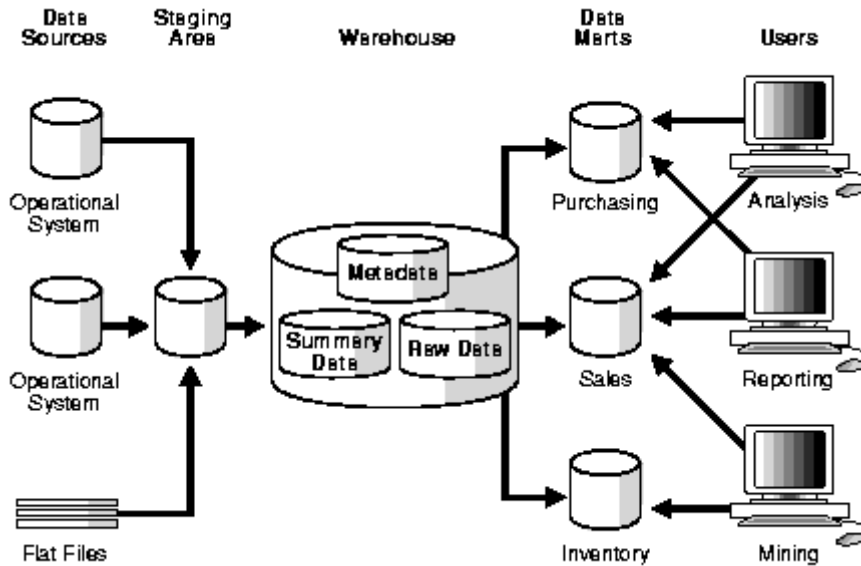


Figure 9: An illustration of data warehouse architecture along with customized data marts (Source: www.oracle.com)

More than transaction processing a data warehouse is a relational database that is designed for query and analysis. Derived from a transaction data usually a data warehouse contains historical data and it separates analysis workload from the workload of transaction further enabling a business or similar projects to consolidate data from diverse sources. Apart from a relational database, environment of data warehouse quite often consists of an ETL solution, an OLAP engine, client analysis tools along with other application that often manage the process of data gathering and delivering to its end users. Further it is possible to classify the data warehouse into following types.

- Enterprise Data Warehouse* – A data warehouse of type generally provide a central database for the entire enterprise for decision support processes.
- ODS (Operational Data Store)* – This data warehouse has a wide scope unlike previous type the stored data is often refreshed in near real time and it is used for regular business activity.

- c. *Data Mart* – It is a subset of a data warehouse that supports a particular function.

Figure 9 shows the technique of transformation of the contents from heterogeneous and multiple data sources that is brought under the umbrella of a common data model further to integrate into a single data warehouse along with the customized data marts. The past years saw the birth of several data warehouses in bioinformatics like ATLAS, BioWarehouse, Columba, DAVID which will be discussed in the next chapter.

2.5 Data visualization

An image is worth a thousand words and it is especially true when describing complex biomolecular interactions (van Iersel et al. 2008). Large amounts of biological data are available from the public databases and domain. It provides great opportunity and also poses a challenge to the life science research and also to bioinformatics community. In order to understand the biological networks and their underlying biological processes, visualization can be a very good analysis tool and it can also act as an intuitive method to explore biological pathways (Ho et al. 2006). Further, to handle and combine large and diverse data sets from various sources and also to gather the information required and to generate new insights from it, the user nevertheless needs strongly user-friendly systems. Since the biochemical knowledge is difficult to conceptualize these systems must have convenient user interface that should analyze the data and also support visualization. Most of the graphical interfaces in case of metabolic networks are designed and they generate only static images of networks and pathways that are expert-curated and replicate the text book view of things (Sirava et al. 2002).

A map of the relationships among the various molecules involved is what required for a system-level understanding of any biological process. With the technologies to detect and predict interactions these maps can be developed. With the help of the map it is possible to study how proteins work in unison forming molecular machines and regulatory pathways. In order to model biological processes and to manipulate and visualize those models there is a need for programs that can generate more dynamic and also more realistic representations of the biological events and structures. An “interactive Cell-TV” is what we need in order to visualize and manipulate models of

cellular events and behaviour. Further it is important, for the iCell-TV, in order to allow navigation of the available information by biologists to operate across seven scales of time and space (Uetz and Finley, 2005). In the past years, several tools and database were developed which integrate and allow the visualization of the molecular data and also generate the biological networks based on different contexts. Some of them are discussed in the next paragraphs.

STRING is a database which is dedicated to protein-protein (**Figure 10**) interactions (physical and functional). The partnerships between the protein functions are the basis of complex cellular phenotypes. Resulting interaction networks formed by the interaction lead to the greater platform for the researchers for modelling, annotation and data reduction. Acting as a meta-database which can map all interaction evidence onto common set of protein of genomes, it weights and also integrates from different sources which includes experimental data sources, public text collections / literatures and from computational prediction methods. Along with other features it is supported with a URL- based programming interface with which it is possible to query the database STRING from other resources (Snel et al. 2000, Jensen et al. 2009).

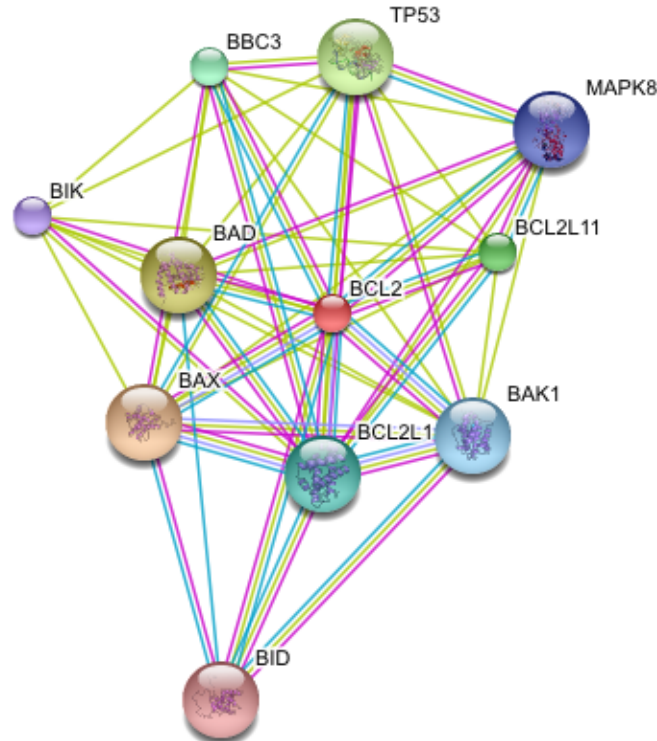


Figure 10: The network result from STRING database shows the protein-protein interactions for Bcl2 related to cardiovascular and other diseases from its integrated data source.

ViSANT is an application that integrates biomolecule interaction data in a very consistent manner. It allows the visual mining of the interaction data in the context of sequence, structure, pathway and other associated annotations. It allows the user to experience different types of manipulation of the data by using the in built functions. It is a integrated software for mining, modelling, analysis and visualization of biological network, further the application is extended to include Gene Ontology (GO) terms for the inference, visualization and analysis of the network. (Hu et al. 2004, 2009)

VANTED is a tool which allows with the related experimental data the visualization and analysis of networks. Via Microsoft-Excel- based form it allows to upload large scale experimental data into the system. Data can be further converted onto network using, either of the three methods, tool itself will draw or it is downloaded from KEGG pathway database or using standard network exchange formats it can be imported. It is possible to present the enzyme, transcript and metabolite data in the context of their underlying networks. Moreover with the support of visualization and navigation it allows to explore the highly enriched data. Supported with the statistical methods comparative analysis of multiple data and their analysis at various developmental stages or genetically lines are possible. It is also possible to generate correlation networks automatically and also to cluster the substances according to their behavioural similarity over time (Junker et al. 2006).

2.6 Pathway Biology

“Systems biology provides a new approach to studying, analyzing, and ultimately controlling biological processes. Biological pathways represent a key sub-system level of organization that seamlessly perform complex information processing and control tasks. The aim of pathway biology is to map and understand the cause-effect relationships and dependencies associated with the complex interactions of biological networks and systems. Drugs that therapeutically modulate the biological processes of disease are often developed with limited knowledge of the underlying complexity of their specific targets. Considering the combinatorial complexity from the outset might help identify potential causal relationships that could lead to a better understanding of the drug-target biology as well as provide new biomarkers for modelling diagnosis

and treatment response in patients” (Ghazal P, 2008). Some of the pathway databases are discussed in the next paragraphs.

2.6.1 Pathway databases

KEGG – Kyoto Encyclopedia of Genes and Genomes is a knowledge data base that links genomic information with functional information of higher order and allowing the systematic analysis of gene functions. It comprises of several databases. And the PATHWAY database stores the graphical representation (**Figure 11**) of higher order functional information. The maps are images and are static and they are curated by experts. KEGG pathway is supported with supplement global maps (Kanehisa et al. 2000, 2008).

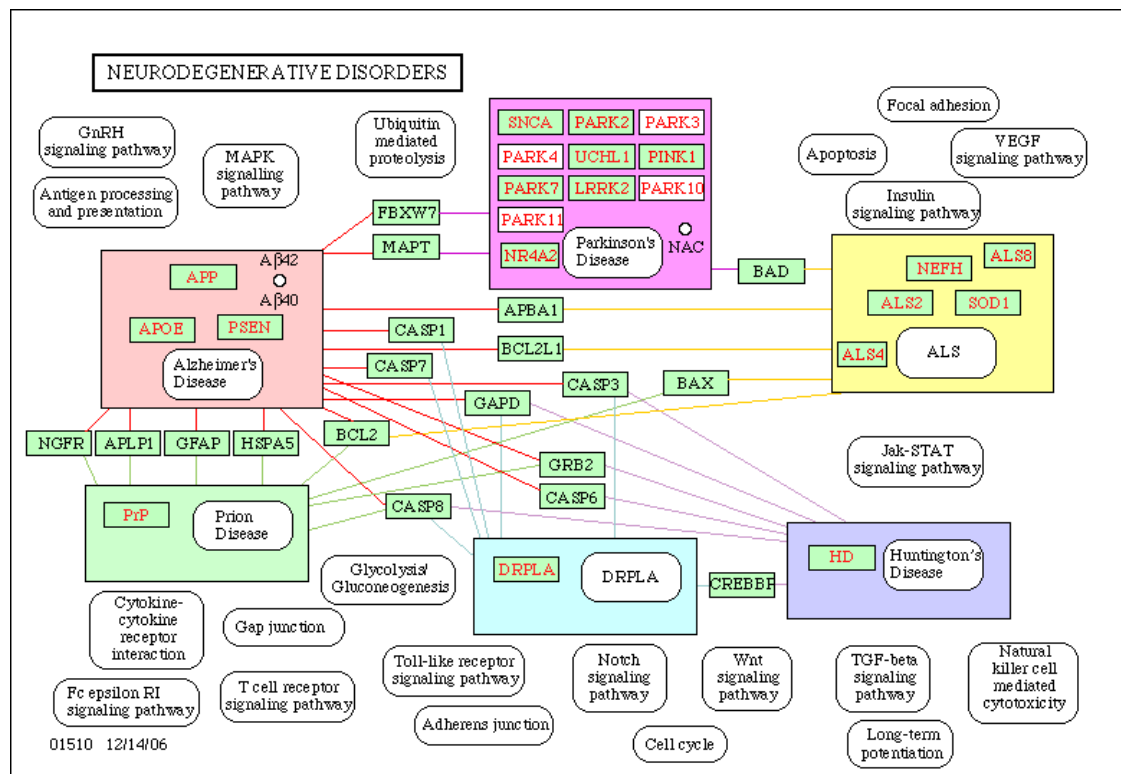


Figure 11: KEGG map showing the various genes including Bcl2 and the pathways associated with neurodegenerative disorders.

Pathway Hunter is a very fast and robust and also a user friendly tool that can analyse the shortest paths in metabolic pathways. It allows the user to perform the analysis for finding the shortest paths using the enzymes involved to build virtual networks (Rahman et al. 2005). Reactome is a curated database intended to provide an integrated view of human biological process and it is peer-reviewed. The very basic unit in Reactome is the reaction and these reactions are grouped together to form the

normal pathways. It allows the possibility of qualitative framework that allows superimposition of quantitative data (Joshi-Tope, 2005).

Pathway Analyst uses the protein sequences to predict the metabolic pathways. Protein sequences are taken as a single FASTA file from a single organism (complete or a partial proteome) and it identifies the sequences which participate in the metabolic pathways available in the server (Pireddu et al. 2005, 2006).

DOQCS – Database of Quantitative Cellular Signaling is a repository of signalling pathway models that is intended to serve the large scale management for signalling chemistry and to the growing field of simulation of signalling networks. The database contains manually created models involving chemical kinetics and it is directed towards the quantitative modelling of signalling pathways. Furthermore the models can be exported to other simulators that are SBML compliant (Sivakumaran et al. 2003).

Even though these and several other databases are available there is always a need for newer databases. This is especially due to the fact, the previously developed databases may lack some property or they would have lesser information or the information which is available from them is not updated regularly making them less preferred. This paves the way for the development of newer databases which has these and some other features incorporated by different research groups in their newly designed databases in order to fulfil the missing links or the information that can enhance the way the rudimentary data or the biochemical knowledge is handled and presented to the community.

2.6.2 Pathway alignment

The term “Align” means to bring into line or line up or arrange in a line and “Alignment” means to position or arrangement or placement. In biology these terms have more significance than just to line up or placement. Here the alignment could involve more than what it states.

If more than two sequences are given we initially

- (a) wish to know or decipher or measure their similarity
- (b) also would like to determine the residue-residue correspondences
- (c) would like to know or observe the patterns of conservations and variability

(d) would like decipher their evolutionary relationships

If all this could be done, then we are in a very good position to mine or fish the databanks for the related sequences. Annotation of genomes which involve assignment of structure and function to as many as genes possible is its major application (Lesk AM, 2005).

To compare two objects, each of which is represented by a collection of elements the most natural way is to try to find element correspondences between the two of them. In a formal representation for example if two objects A and B having elements A_1, A_2, \dots, A_m and B_1, B_2, \dots, B_n , respectively then an *equivalence* is then defined as a set of pairs $E(A, B) = (A_i, B_{j_1}), (A_{i_2}, B_{j_2}), \dots, (A_{i_r}, B_{j_r})$. A equivalence will be termed *alignment* if the elements of A and B are ordered and at the same time if the pairs in $E(A, B)$ are collinear, i.e. if $i_1 < i_2 < \dots < i_r$ and $j_1 < j_2 < \dots < j_r$. (Eidhammer et. al. 2004). Along with the sequence alignment there are interesting research works which focus on the development of algorithms for pathway alignment using the sequence, structure and enzyme based hierarchy approach.

PathBLAST is a tool for network alignment and for searching and comparing protein interaction to identify the conserved pathways and complexes across species by evolution. In a pair of protein interaction paths basically the method searches for a high-scoring alignments between pair of protein interaction paths and finds the occurrence of putative orthologs in the same order in second path as they appear in the first path. Based on the network similarity it allows functional annotation of protein-interaction pathways. The technique can discriminate between true and false positive interactions. With PathBLAST it is possible to graphically view the ranked list of matching paths from the target network along with their overlaps and enables comparative genomics at the network level (Kelly et al. 2004).

MetaPathwayHunter is a pathway alignment tool that can find and report all the similar occurrences based on statistical significance and ranking of the queried pathway or collection of input pathways that is available in the collection. It is an extension of known techniques and is based on a novel graph matching algorithm. It is further supported with visualization method that can project the alignment result of two homologous pathways. It computes the optimal alignment employing a bottom-up dynamic programming approach (Pinter et al. 2005). NetAlign is a web-based tool

that compares a query protein interaction network with the target protein interaction network by a combination of interaction topology and sequence similarity to decipher the conserved network substructures that can derive from a common ancestor and to reveal the topological organization of the network interactions in evolution (Liang et al. 2006). Comparative Pathway Analyzer is a web-based tool that can calculate and display the differences in the content of the metabolic reaction between two sets of organisms. Further it provides hierarchical clustering methods in order to identify the significant groupings. It can also visualize and allow easy comparison of the reaction content of several organisms simultaneously. KEGG is integrated in the system for providing the reaction data and for visualizing the maps. Further, it allows the end users to upload their own annotation data (Oehm et al. 2008).

These systems integrate or utilize the protein information available from different sources and compare using their own algorithm or methods for the prediction, alignment and analysis of pathways. The rudimentary information they analyze can span from the sequence, enzyme to the structure, function to topology and more with a combination of different methods. But still there are some bottlenecks. From the analysis and results of these tools a trail or clue is left about the evolutionary aspects of the proteins and pathways. At the same time leaving some open questions that cannot be answered by them further persuading the need to develop new algorithms that can fair better and also utilize newer information for the pathway prediction, alignment and analysis which could not achieved by these approaches.

2.6.3 Pathway reconstruction

Reconstruction of pathway involves the break down of the pathway into their respective components or reactions and enzymes and then allowing them to view from a perspective of a big or whole network. It involves the mining of the information that are relevant to the pathway and assembling them in a sensible manner to perform different types of analyses. This comprises mining of information from scientific literatures, databases and other resources.

“Even though there is only a single large biological network within any cell and all pathways are to some extent connected, the partition of the entire cellular network into smaller units (e.g. KEGG pathways) is extremely important for understanding biological processes. Biological pathway reconstruction, therefore, is essential for

understanding the biological functions that a newly sequenced genome encodes, and recently, for studying the functionality of a natural environment via metagenomics”. A common method for the reconstruction of biological pathway which is implemented by several automatic biological pathway services and metagenomic sequence annotation starts with the step of identifying the protein function or family such as KO families for the KEGG database in the query sequence. This is followed by a direct mapping of the protein families which are identified onto the pathways. With the adequate knowledge and with the given predicted patchwork of the involved biochemical steps, some judgement or the metric must be applied to decide what pathways do exist actually in the genome or the metagenome which is represented by the sequence. It is more common and straightforward to identify a complete biological pathway in a given dataset provided at least one of the steps that are associated with the pathway is found (Ye and Doak, 2009). Some of the tools used for reconstructing biological pathways are detailed below:

PathwayVoyager utilizes the KEGG online database for the mapping of pathways of partial and whole eukaryotic genomes. Further, it stores the data as local, blast formatted databases by retrieving the user defined subsets of the KEGG database. It marks the ORFs with similarities on pathway maps that are below the end-user defined threshold which is helpful to analyze the whole or partial genomes automatically using NCBI’s BlastP algorithm. It has a straight forward approach and end user does not have to provide substantial resources. Furthermore it utilizes widely accepted resources for the analysis and mapping data and relies on KEGG and Blast algorithm. It is an effective pathway mapping tool for large or confidential data sets despite its uncomplicated approach and can provide evidential data easily during genome analysis (Altermann and Klaenhammer, 2005).

Pathaligner is an alignment and reconstruction tool which uses the enzyme information and hierarchy system for deriving the similarity (**Figure 12**) among the metabolic pathways. It houses the integrated information from different data sources that enables the reconstruction of metabolic pathways from the rudimentary components (Chen and Hofestädt, 2004).

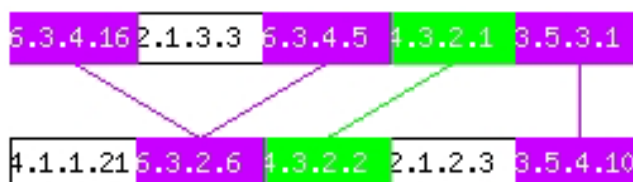


Figure 12: The results from PathAligner shows the reconstructed pathway

ANDVisio (<http://www.pbiosoft.ru/cgi-bin/company/index.cgi>) is a tool for the reconstruction of networks. These are semantic associative networks that allow the user to construct them using their association with gene, protein, metabolite, miRNA, disease, pathway or cellular component. The tool is supported with a search engine that allow to query the underlying database with gene, protein, disease and using other associated object names. Furthermore, it generates interactive visualization network combined with other features.

KAAS (KEGG Automatic Annotation Server) is implemented to assign the KEGG orthology identifiers or K numbers to the genes and genomes. With this it will enable reconstruction of KEGG pathways and BRITE hierarchies. When compared with manually curated methods it has high degree of accuracy and it is based on sequence similarities, bi-directional best information and some other heuristics (Moriya et al. 2007).

2.6.4 Modelling and simulation

“Reductionism is the attempt to explain complex phenomena by defining the functional properties of the individual components that compose multicomponent systems”. It is not possible to understand the complex physiological processes by simply knowing how the parts work in isolation when it is clear that organisms are much more than the sum of their parts. Three main questions are addressed by a systems level characterization of a biological process. (a) What are the parts of the system (i.e. the genes and the proteins they encode)? (b) To know how do the parts work. (c) In order to accomplish a task how do the parts work together (Strange, 2005).

It is usual that the interactions between proteins, genes and complexes can be classified into one or a combination of complex formation, gene activation, inhibition, phosphorylation, equivalent binding and dissociation. It is possible to construct a general framework than can be used to model pathways by translating each of these

interaction types into logic descriptions. Moreover pathways exist in order to process the information within the cell thereby translating these functions in a form. This form does not require the biochemical details to capture the functionality. Rather than classifying based on their location or the proteins and genes involved it is possible to classify the pathways according to their behaviour. This allow to learn how nature builds complex systems and also to know what sub-systems are used as common building blocks further opening the doors to systems-level analysis (Watterson et al. 2008). There is no one particular way to model biological systems (**Figure 13**). For those who think they can model they can start from the bottom and work upwards from the molecular level the ‘bottom-up’ approach; and have to face two problems. The first is computability: hence there won’t be enough computing power to achieve this, despite of having all the necessary information and kinetic equations in place.

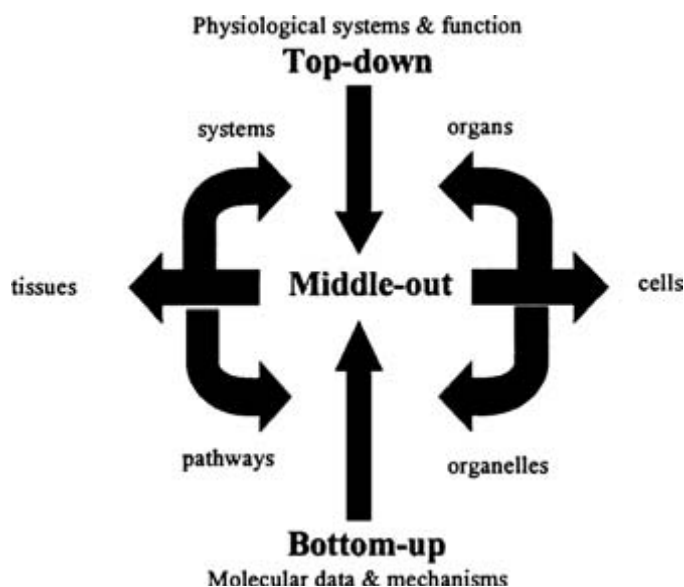


Figure 13: Various strategies for modelling biological processes (Noble, 2003)

The second is the interactions at different levels: one needs to know all the higher levels for a full-scale bottom-up approach, even to characterize fully the lower levels. Hence it is typically impossible only with this approach. Alternative to the ‘bottom-up’ approach is the ‘top-down’ approach where it is intended to start by modelling of high-level physiological systems and then approaching downwards by opening the higher level boxes to typically replace them with progressively more detailed lower-level mechanisms. This approach also suffers problem like the ‘bottom-up’ approach but it is a mirror image. There is no guarantee that one will be able to choose the

correct ones, even though a high-level systems description will be compatible with many different lower-level mechanisms (Noble, 2003).

In systems biology, it requires the integration of experimental data from different sources for the biologically relevant quantitative modeling of physiological processes. Huge amount of experimental data held in numerous public databases is the contribution of the recent developments in high-throughput methodologies enabling the analysis of the transcriptome, proteome, interactome, metabolome and phenome on a previously unprecedented scale. The signals from the extracellular environment and coordinated by intracellular interaction and transcriptional or gene regulatory networks assembled into functional modules modulate the complex systems of cellular and physiological processes. It requires an integrative approach (**Figure 14**) to draw the data from the diverse sources including the data from the public databases, literature, biochemical and kinetic experiments, phenotype studies and also the high-throughput analyses of the genome, transcriptome, proteome, interactome and metabolome in order to understand the cellular processes as interconnected and interdependent systems and also in the context of biological phenomenon (Ng A et al. 2006).

Several tools have been created that allow the modelling and simulation of biological pathways or networks. These tools use the quantitative or kinetic information for their approach. Models and pathways are created either on a common platform or markup language such as SBML that allows the free transformation of the data across different tools which are compliant with the language. In systems biology diverse modelling and simulation methods are being developed and applied. And most of these models can be very easily located spanning the three dimensions of modelling i.e. continuous and discrete, quantitative and qualitative, stochastic and deterministic. All these dimensions are neither entirely dependant nor are exclusive they are hybrid models resulting from the combination of the three aspects. Most important aspect is to distinguish a model to which level it is described whether at the “micro” level or “macro” level or at multiple levels of organization.

“The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biological processes. It's applicable to simulations of metabolism, cell-signaling, and many other topics. SBML has been evolving since mid-2000” (www.sbml.org). SBML is a popular scheme developed by a group and

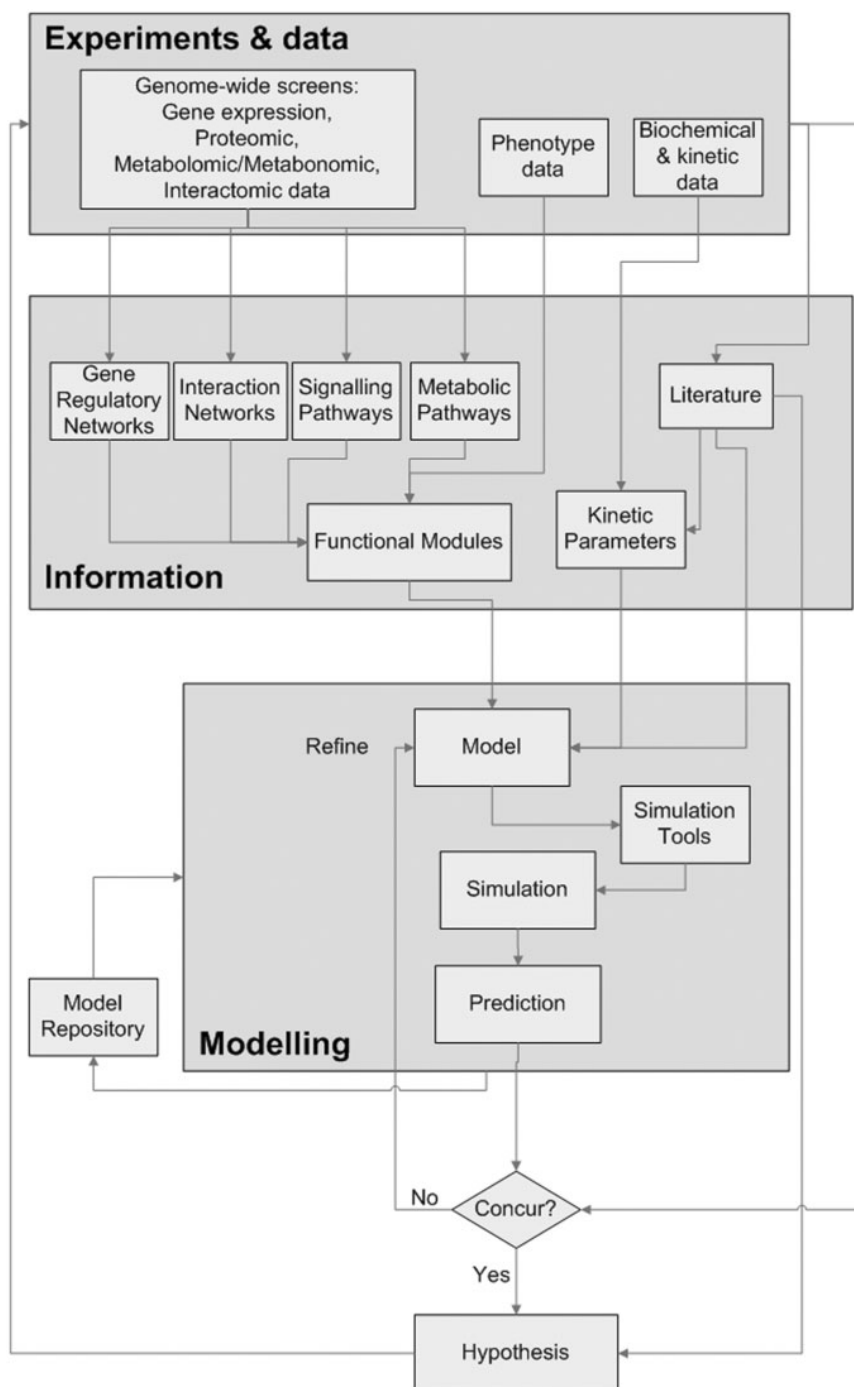


Figure 14: Integrative framework for systems biology (Ng A et al. 2006)

has a community comprising of experts who work on different research domains and are also interested in the development of tools to model metabolic, signalling and many other networks and exchange information across different research groups.

It is possible to break a chemical reaction into a number of conceptual elements such as reactant species, product species, reactions, stoichiometries, rate laws and their parameters. It must be made explicit about the additional components including the

compartments and units of various quantities in order to analyze or simulate a network comprising of reactions. And SBML definition of a model consists of one of these components (Hucka et al. 2003).

Several tools have been developed for biological modelling based on continuous and discrete, quantitative and qualitative, stochastic and deterministic methods like Copasi (Hoops et al. 2006), CellDesigner (Funahashi, 2003) etc. which also allow the import and export of SBML files. Further they also gave rise to the implementation of a common approach and standard along with the requirement for modelling biological networks (Le Novere et al. 2005). An initiative of bringing together the models that are developed across the community as a database (Le Novere et al. 2006) allows the access to the published, peer-reviewed quantitative models of biochemical and cellular systems.

A fundamental question that has been around for several years is always how to model and simulate very complex biological networks. All these years, Petri nets have been captivating the community to solve this key issue. As shown in **Figure 15** Petri net models have been constructed manually using its fundamental components i.e. places, transition and arcs.



Figure 15: Fundamental components of Petri nets.

Source: <http://genome.ib.sci.yamaguchi-u.ac.jp/~pnp/index.html>

In his dissertation in the context of technical systems Carl Adam Petri introduced the basic concepts of Petri nets. An approach was developed by him for describing and studying models, which consist of concurrent, i.e. independent and / or casually dependent components. Many theorems and algorithms have been developed and implemented since this time to analyze Petri net models. Very first applications of Petri net theory to biological systems were published by Reddy et al. and Hofestädt R. Meanwhile, using various classes of Petri nets including the qualitative and quantitative ones, metabolic pathways, signal transduction pathways and gene-

regulatory networks have been modelled and analyzed successfully (Sackmann et al. 2006).

“Petri nets, a graph-oriented formalism, allow the modeling and analysis of systems, which comprise properties such as concurrency and synchronization. A Petri net consists of transitions and places, which are connected by arcs. In the graphical representation, places are drawn as circles, transitions are drawn as thin bars or as rectangles, and arcs are drawn as arrows. The places and transitions are labeled with their names. Places may contain tokens, which are drawn as dots. The vector representing the number of tokens in each place is the state of the Petri net and is referred to as marking. The marking can be changed by the firing of the transitions, which is determined by arcs. The arcs can further be divided into input and output arcs. Generally, arcs may have multiplicities greater than one. A transition is said to be enabled, if all places connected with input arcs contain tokens. An enabled transition may fire by removing a token from each place connected with an input arc and adding a token to each place connected with an output arc. The transitions can be divided into immediate transitions, firing without delay, and timed transitions, firing after a certain delay. A Petri net is a bipartite directed graph, which can be represented graphically. The places contain indistinguishable tokens, which can be fired by the transitions. The vector representing the number of tokens in each place is called the marking of the net” (Hofestädt and Thelen, 1998).

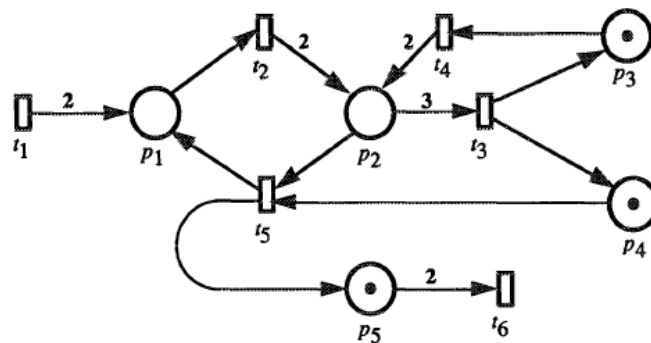


Figure 16: A Petri net graph showing places, graphs and arcs (Reddy et al. 1993)

A Petri net (**Figure 16**) is a graph (PN) consisting of two kinds of nodes i.e. places (p_i) and transitions (t_j). Further these are connected by directed edges called arcs that connect places to transitions and vice versa. Based on the state of the Petri net for each place assigned a non-negative number of tokens. Also, a positive non-zero integer is assigned to each arc which has a weight and it is generally assumed to be

equal to be one if there is no specified arc-weight on the graph. Further each transition i.e. an event is associated with a finite number of input places (pre-conditions) and also output places (post-conditions). A *source transition* which is always enabled is a transition with no input places e.g. t_1 and t_4 in the graph while the others are not enabled. Determined by the weight of the arc an enabled transition can fire and can deposit the tokens in its output places. A transition is a *sink transition* when it has no output places and can fire when it is enabled by consuming the token from its input places. A state of a Petri net (PN) is usually the number of tokens which are present in the individual places and it is usually denoted by M referred as *marking* of the net. M_0 denotes the initial marking of the PN and markings are changed based on the firing of the transitions. And

$$PN = (P, T, E, W, M_0);$$

where

$P = \{P_1, P_2, P_3, \dots, P_m\}$ is a finite set of places

$T = \{t_1, t_2, t_3, \dots, t_n\}$ is a finite set of transitions

E is a subset of or equal to $(P \times T) \cup (T \times P)$ is a set of arcs

$W: E \rightarrow \{1, 2, 3, \dots\}$ is a weight function

$M_0: P \rightarrow \{0, 1, 2, \dots\}$ is the initial marking

P and T being disjoint sets

represents the Petri net in a mathematical way (Reddy et al. 1993). From the following **Table 2** it is possible to typically interpret the transitions and their input and output places (Murata 1989).

Table 2: Typical interpretation of transitions and places

Transition	Input places	Output places
Event	Preconditions	Post conditions
Signal processor	Input signals	Output signals
Computation step	Input data	Output data
Clause in logic	Conditions	Conclusion(s)

Task or job	Resources needed	Resources released
Processor	Buffers	Buffers

There are two main advantages using a Petri-net based approach for biological modelling. First, as a preliminary tool for the analysis of biological pathways many theoretical elements of Petri nets with a mathematical basis are useful. Second, due to the graphical nature of the Petri nets, biologists can easily model and simulate biological systems using the capabilities of the Petri nets (Hardy and Robillard. 2007). Petri net is considered to be the biochemically plausible computational model. For this purpose Petri nets are ideal for modelling parallel and interacting processes. Moreover, a direct analogy can be made very easily between the components and elements in a biochemical reaction networks and Petri net (Mayo, 2005). Cell Illustrator (<http://www.genomicobject.net/member3/index.html>) is a software suite that allows representing and simulating interaction processes in biological pathways based on the notion of hybrid object net and also its basic architecture is based on the biopathway design/representation principle for simulation mechanism along with a XML based technique for visualization. The biological pathway information in concept is described as a graphical representation combined with the explanation about the relation between the concerned biological objects along with the proof of their relationship with the measured / observed quantitative and qualitative data. With Cell Illustrator it is possible to easily represent and simulate the information.

The forthcoming chapters will give a detailed account of the different tools and applications developed. They share some common features with the earlier works of other research groups and are built / based and based on the discussions of previous sections. Also, as described in the previous sections, they are all based on the kernel of this dissertation which is discussed in earlier sections.

Chapter 3

Relational data integration, visualization and network navigation and exploration

3.1 Introduction

Exploration of the enormous amount of data on various molecules and their interactions with their counterparts from day-to-day experiments followed by careful analysis and annotation by experts, has paved the way to create large databases. High and low-throughput and post-genome experiments have further contributed to the growth of the databases. And in life sciences, the integration of heterogeneous data is a growing and well recognized challenge. Further, it is not possible to browse several web pages for the required information within a single database or along several databases. Hence there is a need for a system that allows integration of diverse molecular data. Moreover, instead of just integrating, if it has salient feature which also allow visualization, it will be an advantage, since the human eye is more comfortable with visualization of the facts. These features that are user friendly were not profoundly found among other databases. Here we introduce VINEdb (Visualization Integration Network and Exploration) which can *integrate* the multifarious biological data and at the same time give the user the opportunity to *visualize* the underlying integrated data as a *composite network*. The generated network allows *navigation and exploration* of the molecular data. Furthermore, the data warehouse which stores the information from varied resources has been fortified with a monitor component that allow to *monitor and update* the related data at regular intervals. The KEGG, OMIM, IntAct, GO and UniProt data are integrated in this warehouse. These databases store data in different formats such as tables, maps etc. The data are very diverse in nature and it is also known that they are related to each other in one way or another. Hence the very challenge is to capture, model, integrate and analyze these heterogeneous data in a consistent manner to provide new and deeper insights into complex systems.

* *Parts of the work has been published in 2007, 2008.*

3.1.1 Related works

The presence of numerous and diverse informational resources on genes, enzymes, pathways etc. raises a keen problem of data integration and suitable access of the data. With the ever increasing number of databases and data types, especially in biology, the integration of these heterogeneous data is a challenging problem. The emerging field of integrative bioinformatics provides the essential methods to integrate, manage and analyze the diverse data and allows gaining new insight and a deeper understanding of complex biological systems (Birkland et al. 2006). The difficulty is not only to facilitate the study of heterogeneous data within the biological context but it also more fundamental, how to represent and make the available knowledge accessible (Gopalacharyulu et al. 2005). Moreover, adding valuable information and functions that persuade the user to discover the interesting relations hidden within the data is, in itself, a great challenge (Iragene et al. 2005).

Through database integration methods the information that is present in several databases can be brought together. This integrated data can be further applied for a specific purpose and it can serve research community and it will also allow cooperative information sharing in the field of bioinformatics. In systems biology, the goal of data integration is to unite the diverse information from several databases and data sets, which are obtained both from both high and low throughput experiments, under single data management scheme. The cumulative information can provide greater biological insight than is possible with individual information sources. Large amounts of high dimensional biological data are generated from different throughput experiments and from scientific literature. The data are then carefully analyzed and further annotated by experts for future research and understanding and also stored in different databases. The rapidly growing number of databases and data types poses the challenge of integrating the heterogeneous data types, especially in biology (Baitaluk et al. 2006).

Taking advantage of the data stored in heterogeneous biological data can be difficult and time-consuming for various reasons, which has led to the development of automated systems. The integration of heterogeneous databases is an important issue in biological research resulting in the development of several systems and solutions (Cao et al. 2004). We can still integrate the heterogeneous data according to our needs, thus resulting in a new integrated database. The rapid increase in the volume

and number of data resources drive for providing polymorphic views of the same data often overlap in multiple resources. They are also stored and published in diverse data sources. Each source is distinct in its format. The rich variety of sources available obligates users to visit a myriad of sites to visualize the whole picture of their search for example a protein or gene of their interest. They must also compile the information from journals and then bring in all the details to solve the puzzle (Jayapandian et al. 2007). Hence the graphical representations of the facts are comparatively easier to understand than if they were presented as raw data. This is especially true for large datasets or complex situations (Uetz et al. 2002).

It is usual in protein-protein interactions the nodes represent proteins and the edges represent the protein-protein interactions when they are visualized as a graph. And when dealing with the smaller number of nodes and edges the visualization of a graph is straightforward approach. But in practice, this is not possible as protein-protein interactions consists of several thousand nodes and for many graph drawing tools, this is a limiting step due to the fact they produce cluttered drawings with too many edge crossings or static drawings, that cannot be modified or hard to modify. Further, for user interaction they are too slow and also require the data in a specified format instead of using the data directly from the databases. The protein-protein interaction network should convey its message clearly and quickly since the ultimate value of visualizing interaction on a graph depends on its readability (Ju and Han, 2003).

The rudimentary data is provided from the distributed, heterogeneous data sources to the end user in a homogenous way by the text indexing systems and multi/federated database systems. With the increasing amount of life science data, there is a change in the trend from the pure data management to comprehensive and complex data analysis in bioinformatics. The data warehouse systems provide a universal data schema and it allows periodical loading of the data into a central repository. With the help of data warehousing concept the major limitations of the distributed database systems such as inconsistency of data and consumption of time or query incompleteness caused by the server restrictions are overcome (Fischer et al. 2006). And in molecular biology the idea of data integration is not new one. Several projects have focussed on the challenging problem of interoperability among the different biological databases. In the early nineties it was first addressed by P. Karp about biological database integration (Karp 1995). And several integration approaches and sources have been

developed for the integration of the diverse molecular biological data. These are based on different data integration techniques, e.g. text indexing systems (e.g. SRS, BioRS), multi database and federated database systems (e.g. DiscoveryLink, BioKleisli/K2), and data warehouses (e.g. Atlas, BioWarehouse). IBM developed the DiscoveryLink (Haas et al. 2001) system to access the heterogeneous data source by one single SQL query. As it is based on the federated database technique, it requires the development of a global data scheme. With the help of wrappers and views DiscoveryLink access its original data. The SQL, available in read only format is supported as query language. As such the system is now part of IBM's Websphere Information Integrator.

The data warehouse approach has several advantages unlike the database solutions in which either the query is not proper or the response is not complete, thus allowing complex querying to be faster and also less time-consuming. In bioinformatics there are several examples of data warehouse serving the community at various levels with diverse data. As such these projects can be roughly separated into two groups: (a) general software infrastructure for further customization within the bioinformatics applications (e.g. Atlas, BioWarehouse) and (b) project-oriented data warehouse (e.g. Systomonas, Columba) implementation that are developed to answer particular biological questions (Hariharaputran et al. 2007).

3.1.2 Bio-data warehousing approach

The huge quantity of information generated in life sciences is dispersed in many databases and repositories. In spite of the accessibility, there is a great demand for the methods that will be able to gather and display the distributed data in a friendly and standardized manner (Cases et al.2007). In Atlas, the data warehouse, the biological sequences and information of molecular interactions, homology, functional annotation of genes and the ontologies of biology are stored locally and integrated with the goal of providing the data and as well as a software infrastructure for enabling bioinformatics research and development. Moreover Atlas stores and integrates local instances of Biomolecular Interaction Network Database (BIND), Database of Interacting Proteins (DIP), IntAct, NCBI Taxonomy, Gene Ontology (GO), Molecular Interactions Database (MINT), GenBank, RefSeq, UniProt, Human Protein Reference Database (HPRD), Online Mendelian Inheritance in Man (OMIM), Entrez Gene, LocusLink and HomoloGene. It offers end-users flexible, easy, integrated access to

this data with the retrieval APIs and toolbox applications that are critical components (Shah et al. 2005).

BioWarehouse is an open-source toolkit for constructing bioinformatics databases using different database management systems such as MySQL and Oracle. The component databases are integrated into a common representational framework under a single database management system (DBMS). It facilitates different database integration tasks such as data mining and comparative analysis along with enabling multi-database queries using the Structured Query Language (SQL). In addition to UniProt, GenBank, NCBI Taxonomy, and CMR databases and the Gene Ontology, it supports and integrates the pathway-centric set of databases which includes ENZYME, KEGG and BioCyc. Written in JAVA and C languages the loader tools help to parse and load the databases in the database scheme. The loaders reduce the semantic heterogeneity by applying a degree of semantic normalization to the source data. A variety of bioinformatic data types are supported by the scheme such as metabolic pathways, proteins, genes, nucleic acid sequences, chemical organisms, organism taxonomies, and controlled vocabularies, compounds, biochemical reactions, features on protein and nucleic-acid sequences. BioWarehouse making a significant progress is another approach to answer some of the complex queries which adds value to the data warehouse approach and data integration methods in bioinformatics (Lee et al. 2006). Neither Atlas nor BioWarehouse is platform-independent, as each of them is implemented with different programming language and requires lot a time for installation.

Systomonas is an integrated informatics platform that has been developed for systems biology approach and for the analysis of the biology of pseudomonas in biotechnology and in infection. It stores the prediction information of cellular processes such as gene regulatory networks along with the in-house experimental metabolome, proteome and transcriptome data (Choi et al. 2007). Columba integrates several databases like PDB, KEGG, Swiss-Prot, CATH, SCOP, the Gene Ontology, and ENZYME. It allows the end user to search the database using either a keyword or using data source-specific web forms. As such the users can quickly select and download PDB entries for example participating in a particular pathway or have been classified under a certain CATH architecture or having certain molecular function and have been annotated in the Gene Ontology and the structures have a particular

resolution under a different threshold. End users are provided both with machine-readable markup language as well as human readable format results. Furthermore, it allows the interactive viewing of the structures on the web (Trissl et al. 2005).

Both Systomonas and Columba are available via the internet and they are supported with web-based graphical user interface that can be accessed with any web browser. At the same it is hard to judge the up-to-dateness or the newness of the integrated data. In Reactome which is a knowledge base for biological process the basic unit is a reaction and the reactions that are grouped into casual chains for the pathways. By this it is possible to infer the reaction equivalence both in human and non-human species. Furthermore it is supported with a “Skypainter” tool that colorizes the chosen genes or compounds participating in the given reaction (Tope et al. 2005).

Table 3: A comparative table showing the salient features of different data warehouses

	Atlas	BioWarehouse	Columba	Systomonas
Institute	British Columbia University, Canada	Stanford Research Institute, USA	Humboldt University Berlin, Germany	Technical University, Braunschweig, Germany
Objective	Supply of data and software infrastructure	Development of user specific data warehouse instances	Data integration about protein structure and function	Database about molecular networks in Pseudomonas
Integration	Close integration ready made relational schemas	Close integration ready made relational schemas	Loose integration, multi-dimensional model	Unknown
DBMS	MySQL	MySQL and Oracle	PostgreSQL	PostgreSQL
Language	Java, C++,Perl	Java, C	Python, Perl	PHP
Architecture	Software infrastructure	Software infrastructure	Web application	Web application
Complexity	Time consuming local installation	Time consuming local installation	Only web browser	Only web browser
Web Interface	Exiting example, but not available	Example “Public House” available	Complete	Complete
Platform independent	Only Unix based	Only Linux based	Yes	Yes
Update	Manually	Manually	Old data	Unkown
License	Source code available (GNU)	Source code available (Mozilla PL)	Parser available on request, web application free	Web applicaton free

In **Table 3** the salient features of different data warehouses are compared. Based on these and previous considerations, we designed a platform-independent data warehouse system, VINEdb (**Figure 17**) that integrates heterogeneous data sources into a local database and provides a comprehensible updating strategy to ensure a maximum transparency and up-to-dateness of the integrated data. It is a web-based data warehouse that supports scientists exploring integrated life science data from popular molecular biology data sources

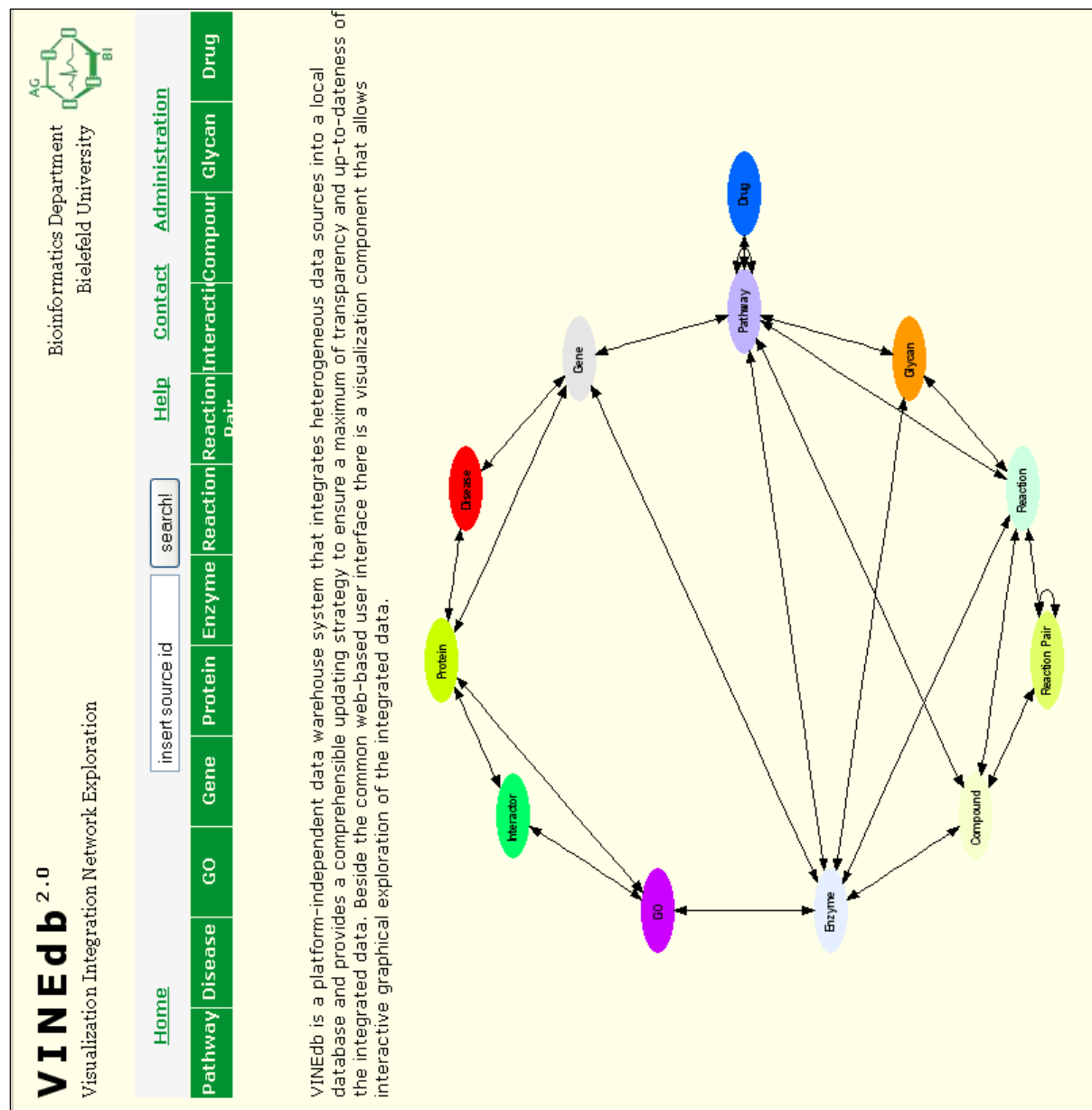
3.2 VINEdb - Visualization Integration Network Exploration

To know and show what happens in a cell a long list of interactions or a protein pairs table will not be sufficient. For this purpose, a graphical representation of the facts makes it easier to understand the complex situations or biological complexity. Moreover, the use of graphics or graphical images suits the human preference for visual perception (Uetz et al. 2002).

3.2.1 Concept

The key idea in our work is not only to develop a data warehouse that integrates and manages diverse data, but also to emphasize the visualization of the integrated data. The advantage of visualization method is it provides the user with more information about the rudimentary data in a comprehensible manner. VINEdb combines the methods of interactive visualization and graphical representation of the facts to make things easier. Moreover creating network graphs at run-time that are dynamic and as well as interactive will be more advantageous giving the opportunity and accessibility to the user to explore and navigate through the interconnected domain information in the system. A menu at the top panel helps to navigate through the contents of the data warehouse such as pathways, protein, enzymes, disease, drug, etc. In VINEdb the key idea is to integrate the diverse information and represent the relation as one or more interactive network graphical images. For example, a gene and its relation to other proteins, enzymes, disease, drugs, compounds, interactors and pathways can be illustrated with one or two networks/graphs further providing the associated information within the data warehouse along with the links and details to the source. By generating a composite network based on the user's interest from the integrated data, this method is effective in determining the relationships between the diverse data. Furthermore the extendable open source data warehouse architecture of this

system enables platform-independent use of the web application and the underlying infrastructure.



VINEdb is a platform-independent data warehouse system that integrates heterogeneous data sources into a local database and provides a comprehensive updating strategy to ensure a maximum of transparency and up-to-dateness of the integrated data. Beside the common web-based user interface there is a visualization component that allows interactive graphical exploration of the integrated data.

Figure 17: The interactive domain information system as in VINEdb. The graphical image allows navigation and exploration of the interconnected data.

3.2.2 Architecture

VINEdb has four-layer system architecture as illustrated in **Figure 18**. The data sources OMIM, KEGG, UniProt, IntAct and GO form the source layer which is the basis of the system and contain the original data. It is provided by parseable flat files by most of the databases, whereas Gene Ontology (GO) is supplied via a SQL dump file. Only a small extract of data is integrated from PDB and Enzyme

(<http://www.chem.qmul.ac.uk/iubmb/enzyme/>). The different external data sources are controlled by the automated monitor component of the integration layer to ensure a high degree of transparency and maintain up-to-dateness. It recognizes changes of the original source and starts a download of the changed files if necessary. The files, when downloaded successfully to the local file system, will activate the parser to start the ETL process. ETL is a typical data warehouse process and marks the successive steps of extraction, transformation and loading of data. That means the data is extracted from the original data exchange format, transformed to the target data

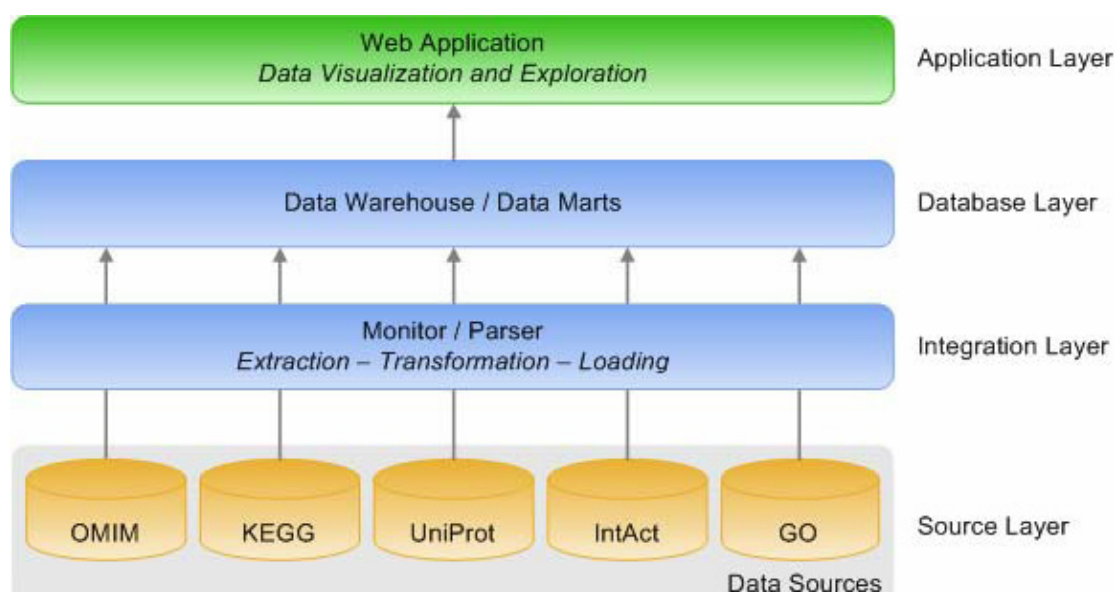


Figure 18: The system architecture of VINEdb illustrates the flow of data upstream from the original heterogeneous data sources source layer to the web application layer in between checked by the monitor component for data up-to-dateness and then integrated into the database layer.

schema and loaded to the data warehouse. Afterwards, the data can be queried and accessed by the web application. Thereby the export schemata of the data sources are loosely coupled by mapping tables that establish relationships between data entities from different biological aspects. As such this approach simplifies the maintenance of the data schema and also the integration of new data sources. And based on the data that is loaded into the data warehouse of the database layers, smaller data marts can be easily constructed for specific analysis and applications. The end-user can interact with the system by a web-based graphical user interface -the web application. As previously described, this web application allows users homogenous access to the integrated data and supports its exploration by interactive visualization of

relationships between data entities. The web pages of the system are accessible with any common browser. Each of these data entities is further supported by detailed information and also further referred to its or their original data source. Also it is supported with a search engine which facilitates and allows the end users to interact at different levels to find the information of their interest.

3.2.3 Implementation

The core of this data warehouse infrastructure is implemented completely in Java to ensure platform independence of the operating system. Thus it can be used separately from the web interface to integrate several life science data sources into relational databases for individual research. This task is supported by a collection of ready-to-use parsers for standard life science data sources e.g. Brenda, EMBL, ENZYME, GO, iProClass, KEGG, OMIM, PubChem, Taxonomy, UniProt, SCOP and CATH. Currently, the preferred database management system for the integrated data is MySQL, but an additional persistence layer is under development to enable the use of further relational database management systems e.g. Oracle or PostgreSQL. Once a release candidate of the software is finished, it will be available on SourceForge (<http://sourceforge.net/projects/biodwh/>). **Table 4** shows the list of biological objects integrated into the system and their original public life science database.

Table 4: List of biological objects and corresponding public life science database integrated into VINEDb.

Objects	Source databases	Property	External links to the source databases
GO-Term	GO	Gene Ontology-Term	http://www.geneontology.org/
Interaction	IntAct	Interaction	http://www.ebi.ac.uk/intact/
PDB	PDB	Protein structure information	http://www.rcsb.org/pdb/
Gene	KEGG	Gene information	http://www.genome.jp/kegg/genes.html
Pathway	KEGG	Pathway maps	http://www.genome.jp/kegg/pathway.html
Enzyme	KEGG	Enzyme	

		nomenclature	http://www.genome.jp/kegg/ligand.html
Reaction	KEGG	Chemical reaction	http://www.genome.jp/kegg/ligand.html
Reaction Pair	KEGG	Reactant pair alignments	http://www.genome.jp/kegg/ligand.html
Compound	KEGG	Chemical compound structures	http://www.genome.jp/kegg/ligand.html
Glycan	KEGG	Glycan structures	http://www.genome.jp/kegg/ligand.html
Drug	KEGG	Drug structures	http://www.genome.jp/kegg/ligand.html
Disease	OMIM	Disease	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
Protein	UniProt	Protein	http://www.expasy.uniprot.org/

Following is the technical representation (**Figure 19**) that depicts how the biomedical data is being integrated into VINEdb. OMIM, UNIPROT, GO, KEGG and IntAct are the data sources for building the data warehouse. The data sources are monitored regularly for their updates. Further, the parsers written for the transfer will assist for converting the source files supplied by the databases and then for transferring and integrating the data into VINEdb. Each of these parsers scan the source files and then dumps the newly converted tailor made data to the data marts. Data marts containing the newly converted data will supply them to the main data warehouse, VINEdb. The web application that is connected to the data warehouse allows the end user interaction. The interrelationship between the source databases is shown in **Figure 20** provide the information how the different domain and entities presented to the end user is connected in the underlying database. It shows the domain data or the information being stored in the form of tables and further each of these tables are cross referenced during the search by the end user. The result generated after the search will be presented in the web browser in the form of a graph or interactive image which allows further navigation.

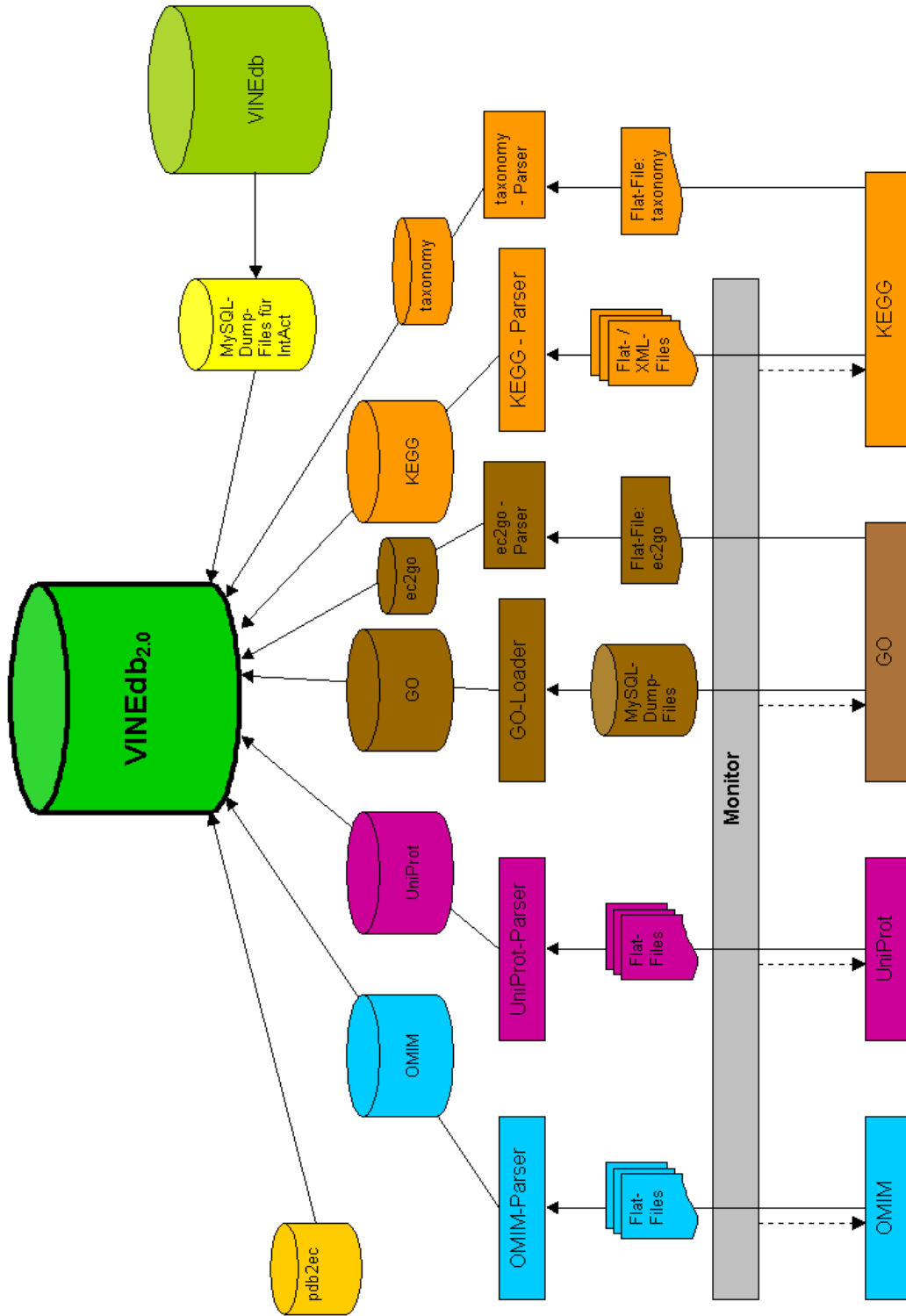


Figure 19: A graphical representation of the integrated biomedical data in VINEdb (Fietz, 2007)

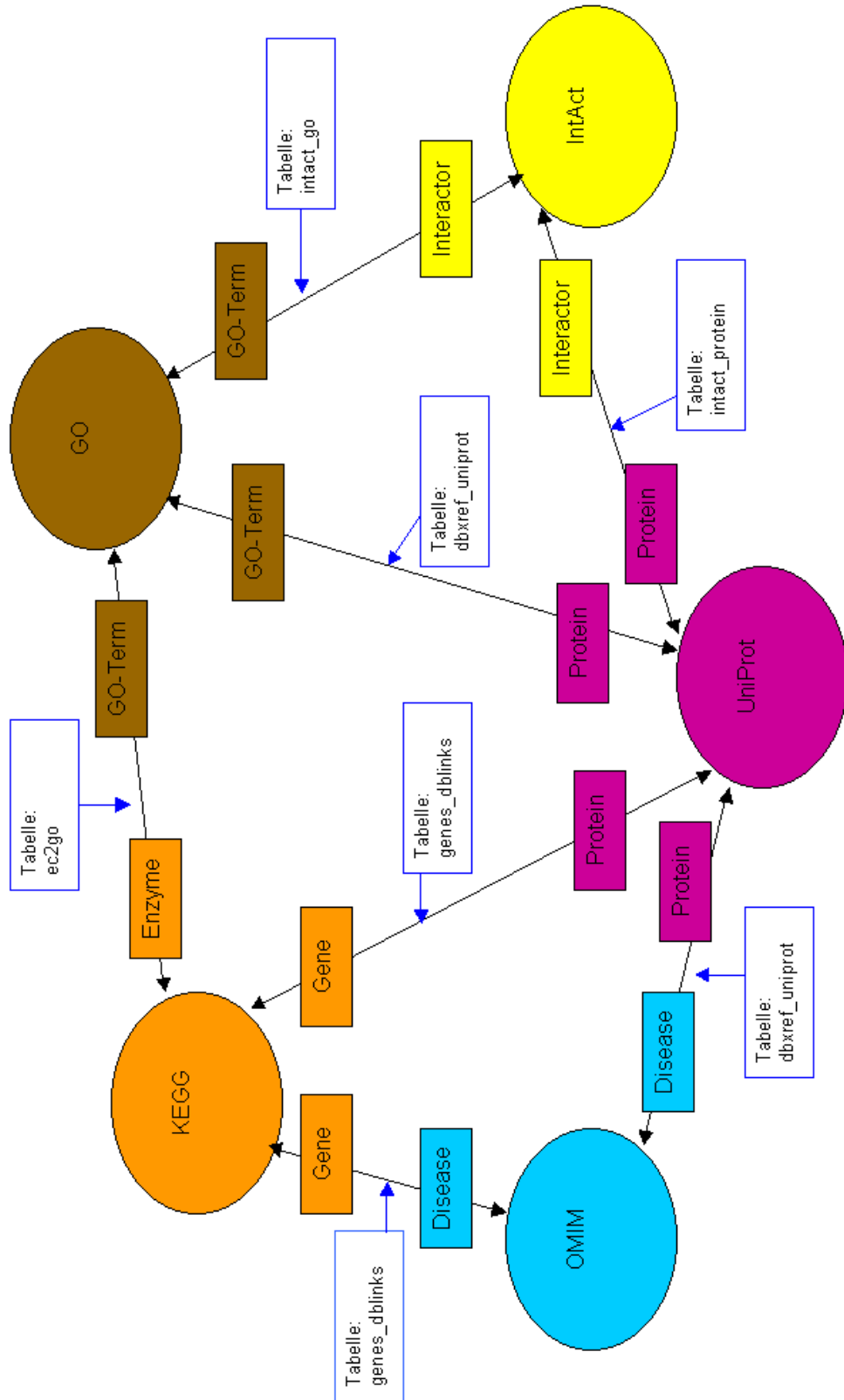


Figure 20: The graphical representation of the relationship between the different data sources which are integrated into the system showing how they are connected with each other. (Fietz, 2007)

The web-based graphical user interface of VINEdb is implemented with JavaServer Pages and runs on an Apache Tomcat web server. Based on user activity, it carries out different search, preparation and presentation functions by connecting to the database and generating HTML pages. The graph visualization software Graphviz (Gansner et al.1999) is used to dynamically create the graphical representations of the relationships between the entities of the data warehouse. Graphviz (<http://www.graphviz.org/>) is controlled by the DOT language that provides syntax to describe graphs, nodes and edges with additional layout preferences. Thus, a DOT file is generated according to the entities selected by the user and their relationships it is given to the Graphviz software that produces a PNG image file with the graph visualization. Afterwards, this image is embedded in the HTML pages and displayed by the web browser. The web-based graphical interface enables the end user to interact with the system and gives users equal access to the integrated data and supports the visualization of relationships between data entities. The web pages of the system can be accessed with any common browser.

Having considered the features of other data warehouses which are explained earlier (Table 4) the comparative **Table 5** lists and compares the features of VINEdb along with the other data warehouses. VINEdb is developed and implemented in-house using data integration methods and for providing easy access to versatile data from multiple databases and also to provide the related molecular information in an efficient manner to the researchers. This is useful for biologists who could get the information as a capsulated or composite network and will be able to avoid browsing through several pages to get the same information. This was not possible with other data warehouses and sometimes they are not available free or had some other technical problems which prevented using them.

Table 5: A comparative table showing the features of different data warehouses with VINEdb

	Atlas	BioWarehou se	Columba	Systemonas	VINEdb
Institute	British Columbia University, Canada	Stanford Research Institute, USA	Humboldt University Berlin, Germany	Technical University, Braunschweig, Germany	Bielefeld University, Germany
Objective	Supply of data and software	Development of user specific data	Data integration about protein	Database about molecular	Integrated DataWarehose use

	infrastructure	warehouse instances	structure and function	networks in Pseudomonas	
Integration	Close integration ready made relational schemas	Close integration ready made relational schemas	Loose integration, multi-dimensional model	Unknown	Loose integration
DBMS	MySQL	MySQL and Oracle	PostgreSQL	PostgreSQL	MySQL
Language	Java, C++,Perl	Java, C	Python, Perl	PHP	Java, JSP
Architecture	Software infrastructure	Software infrastructure	Web application	Web application	Web application
Complexity	Time consuming local installation	Time consuming local installation	Only web browser	Only web browser	Easy
Web Interface	Exiting example, but not available	Example "Public House" available	Complete	Complete	Complete
Platform independent	Only Unix based	Only Linux based	Yes	Yes	Yes
Update	Manually	Manually	Old data	Unkown	Regular
License	Source code available (GNU)	Source code available (Mozilla PL)	Parser available on request, web application free	Web applicaton free	Freely available

3.2.3 Application

The huge data from different sources like KEGG, GO, UNIPROT and IntAct are quite difficult to handle. At the same time it is time consuming to browse across the web pages to bring in the needed information. Therefore, there is a need for a representative or a consolidated image that can project or give an overview of the background information of the integrated sources (Hariharaputran et al. 2007). The protein and other related information are stored in different databases across the globe deposited by different research groups and are heterogeneous in nature. As an input and to demonstrate the salient features of VINEdb that can be useful for biologists and other experts the proteins/genes which are used as inputs are generated by different experiments conducted by the biologists/medical experts in the EU CardioWorkbench project. Among them Bcl2, CASP3, Rac1 along with some other proteins have been found to have a significant role in cardiovascular diseases and it has been proved by their experiments. These proteins in turn are also connected to various diseases,

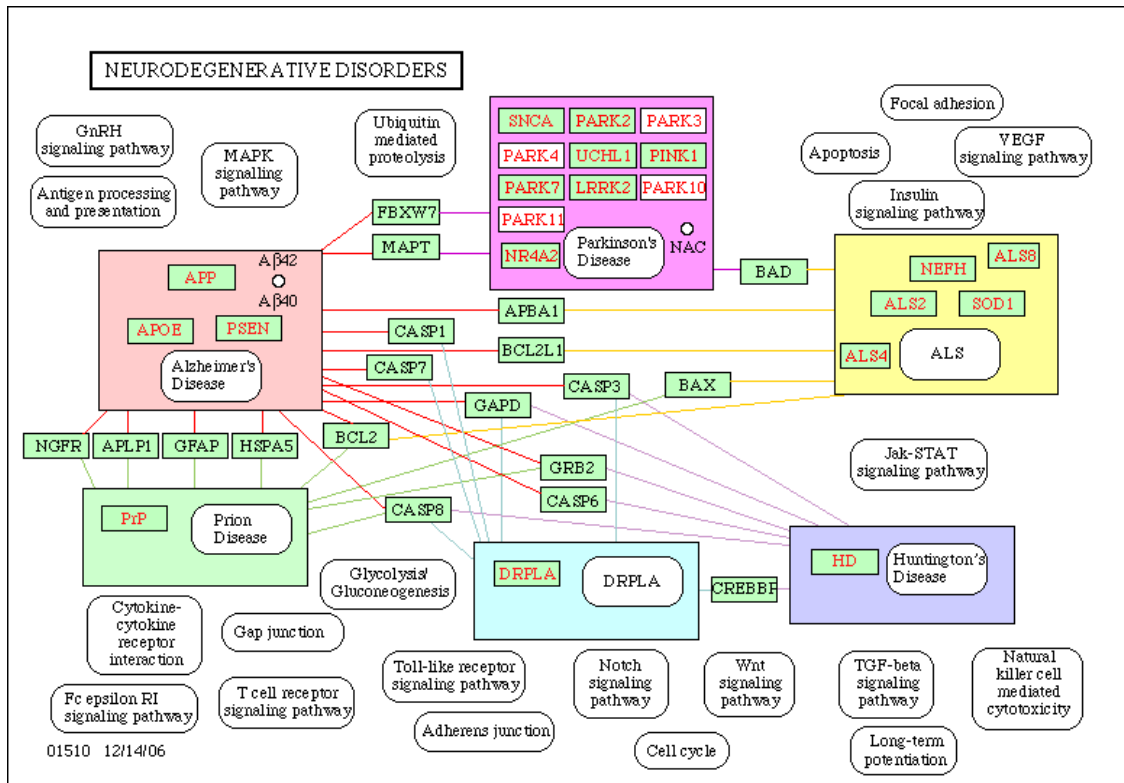


Figure 21: KEGG map showing the role of Bcl2 in neurodegenerative diseases with different pathways.

pathways, enzymes etc. By knowing their relationship with other rudimentary data or information it is possible to form a hypothesis that can be utilized for designing drugs for cardiovascular diseases and help in their future experiments. For e.g. Bcl2 a protein is involved in cardiac, neurodegenerative and other diseases as shown in the KEGG map (**Figure 21**). And at the same time its related information is also stored in different databases like UniProt and KEGG (**Figure 22**) which in turn stores the protein, pathway and gene information. But it is not possible to browse several databases and web pages to retrieve the information what we needed about the above said proteins/genes. Without consuming much time if it is possible to generate a network of information it will be the most sought way for any biologist/researcher who work with these proteins. And with VINEdb this is possible as it allows easy interaction and quicker generation of related information as a composite network, also less time consuming.

UniProtKB/Swiss-Prot entry P10415

[Entry info] [Name and origin] [References]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual.

Entry information

Entry name	BCL2_HUMAN
Primary accession number	P10415
Secondary accession numbers	P10416 Q13842 Q16197
Integrated into Swiss-Prot on	July 1, 1989
Sequence was last modified on	April 1, 1993 (Sequence version 2)
Annotations were last modified on	

Name and origin of the protein

Protein name	BCL2
Synonyms	
Gene name	BCL2
From	Homo sapiens (human): 596
Taxonomy	
Protein existence	

References

[1] NUCLEOTIDE SEQUENCE [Mature] from Homo sapiens (human). PubMed=3523487 [NCBI, Exon] Tsujimoto Y., Croce C.M., "Analysis of the structure, transcription, and localization of the BCL-2 gene." Proc. Natl. Acad. Sci. U.S.A. 86:1398-1402 (1989).

[2] SEQUENCE REVISION TO SWISS-PROT P10415. DOI=10.1093/nar/20.16.4187

KEGG Homo sapiens (human): 596

Entry	596	CDS	H.sapiens
Gene name	BCL2		
Definition	B-cell CLL/lymphoma 2		
Orthology	KO: K02161 apoptosis regulator BCL-2		
Pathway	PATH: hsa01510 Neurodegenerative Diseases PATH: hsa04210 Apoptosis PATH: hsa04510 Focal adhesion PATH: hsa05030 Amyotrophic lateral sclerosis (ALS) PATH: hsa05060 Prion disease PATH: hsa05210 Colorectal cancer PATH: hsa05215 Prostate cancer PATH: hsa05222 Small cell lung cancer		
Class	BRTE hierarchy		
SSDB	<input type="button" value="Ortholog"/> <input type="button" value="Paralog"/> <input type="button" value="Gene cluster"/>		
Motif	Pfam: BH4 Bcl-2 PROSITE: BH1 BH2 BH3 BH4_1 BCL2_FAMILY BH4_2 <input type="button" value="Motif"/>		

Figure 22: Protein and gene information of Bcl2 human from Uniprot and KEGG databases

Through the web pages any user can interact with the system home page or search engine available to mine the information from the integrated data. The end user can navigate through the different domains equipped with individual search engines which allow them to type in the forms provided for the proteins, genes, pathways, etc. for more details. Every form is supported with appropriate example at each level to help the user to know the type of input required. The input can range from name, id, definition, to symptom, E.C. number, etc. according to the query. Furthermore, a filter is provided to restrict the search to a particular organism. **Table 6** shows different biological objects and possible search parameters in VINEdb. The gene Bcl2 involved in cardiac and other diseases in human is searched in the search engines and result or the output of VINEdb is shown in the next paragraphs.

Table 6: List of biological objects for search and allowed parameters in VINEdb

Entity	Search parameters
Compound	Compound id, Compound name

Disease	MIM number, MIM title
Drug	Drug id, Drug name, Drug activity
Enzyme	Enzyme number, Enzyme name, Enzyme class
Gene	Gene id, Gene name, Gene definition
Glycan	Glycan id, Glycan name
GO	GO id, GO name
Interaction	Interaction id, Interactor id, Interaction name, Experiment id
Pathway	Pathway id, Pathway name, KEGG organism
Protein	UniProt id, UniProt accession number, UniProt gene name, Filter by organism
Reaction	Reaction id, Reaction name
Reaction Pair	Reaction pair id, Reaction pair name

When the user types in the search term and initiates the search, the system will look for the information and the output will be a web page of entity information. The web page provides information and also links to the graph and other data such as pathway, disease, map, references, etc. There is also a link to the source of information in output pages. The graph is dynamic and is predominantly interactive, allowing the user exploration and navigation for further information. The images representing the interaction can be saved in PNG format. The search for the protein Bcl2 in VINEdb will result in an entity information page with a link to the graph. **Figure 23** shows screenshots of the graphical user interface example. The network graph is a representative or consolidated image of the related domain information at the first level and further allows the user to interact, explore and navigate through each relation like the gene, disease, GO and other associated terms with simple browsing methods.

The details of several pathways that share some information and are connected to CASP3, the apoptosis related Cystine peptidase along with the enzyme, disease and other related protein information is shown in **Figure 24**. Among the pathways that are listed in the image are MAPK, apoptosis pathways, neurodegenerative disorders map along with others along with related information.

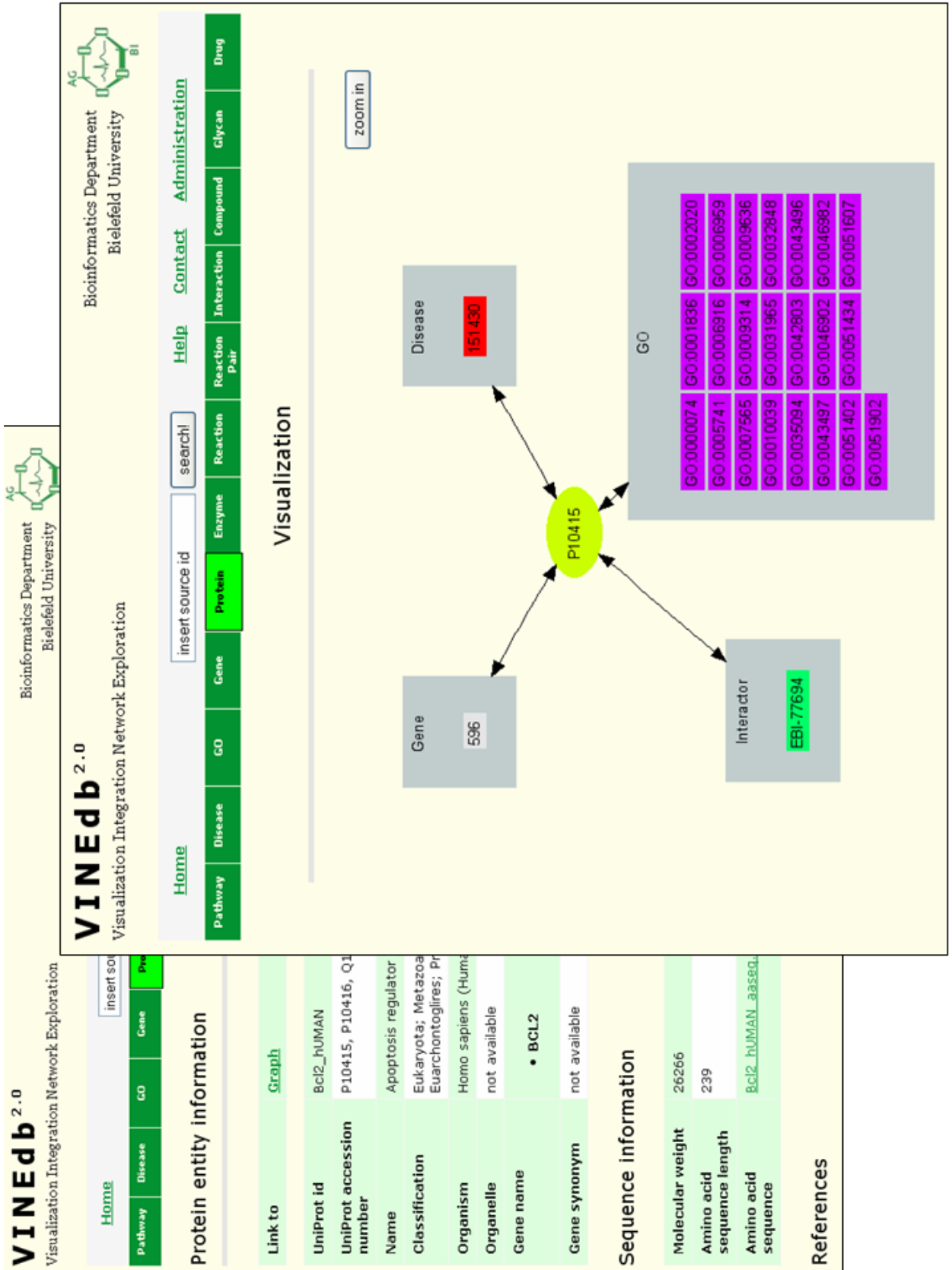


Figure 23: This image illustrates the results of the search showing related entity information for apoptosis regulator Bcl-2 in human and the interactive graph

generated from VINEdb. The links and the graph allow further navigation and exploration.

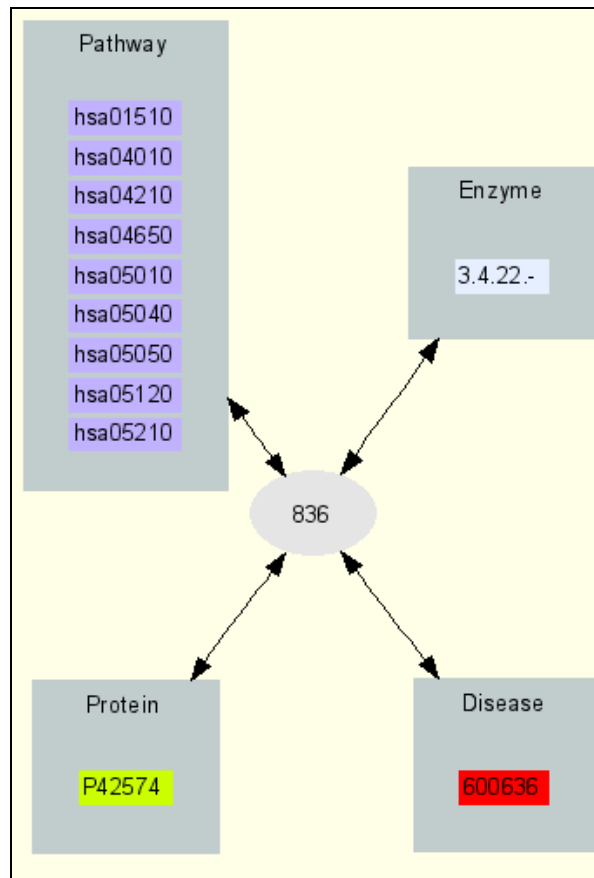


Figure 24: CASP3 (Gene id 836) is connected to several pathways including MAPK (hsa04010), Apoptosis (hsa04210) and Neurodegenerative disorders (hsa01510) along with protein and disease information available within the data warehouse.

The pathways in turn are related to cardiac related diseases and genes like Bcl-2 and Rac1 which is possible to explore through the network navigation and exploration approach. It is possible to get more information by browsing either of the associated information in the interactive network image that will further leads to their related entity information and more details.

3.3 Summary

One growing and well-recognized challenge in life sciences is the integration of heterogeneous data. The availability of heterogeneous data in different databases is well-known and is stored in different formats. This makes it hard to comprehend the interactions, relationships and underlying complexity between the databases. But the same data, when integrated and presented in an effective way, will allow users to gain

more knowledge about a protein/gene and its interacting partners and to know about phenotype/genotype relationships.

To represent and make it accessible the available knowledge is more a fundamental problem, apart from the difficulty to aid the study of such multifarious data within the biological context. VINEdb is a new approach and a blend of relational visualization and data integration methods that allows interactive exploration and navigation of integrated life science data. Fortified with a monitor component that helps to keep abreast of the integrated data, it paves the way for presenting complex biological data in a very user-comprehensible manner. Such is not possible using raw or long lists of interaction data or with a protein table. Furthermore, VINEdb among other functions enables interactive visualization, network navigation and graphical representation making it possible to present in one or more images the different domain information. Thus, with a consolidated image the users will be able to explore and understand the complex nature of the molecular interactions involved, with ease. The generated network from VINEdb among different relations also focuses on the relations between proteins and pathways. These pathways share some information with each other and this information can be useful to compare the pathways. If it is possible to compare the relationship among pathways then it is possible to find the homology among disease pathways too. This further can help the scientists in the CardioWorkBench and other projects to use the knowledge in their findings.

VINEdb is available at <http://agbi.techfak.uni-bielefeld.de/VINEdb/>.

3.4 Outlook

The information that pertains to the genes, pathways, proteins, diseases etc are widespread and new data are also generated each day. Hence it is possible to bring in more information into VINEdb. The data warehouse can be expanded further in order to accommodate more information. It is also possible to make changes with the way how the data is presented by adopting new visualization methods. Moreover the integrated information can be adapted or used for new concepts that can give a newer dimension of the integrated data.

In the previous chapter a method to mine and display the information or the data from the data warehouse is proposed. The information pertained to the genes, proteins, diseases, pathways etc. The forthcoming chapter introduces a method that can align the biological pathways or networks using integrated information system that comprises of protein structural information along with the enzyme information. The pathways may be connected to different diseases and can also share similar proteins in different organisms. Therefore a system that can significantly utilize the rudimentary biochemical knowledge of the proteins involved and can also decipher the evolutionary relationship of the proteins and pathways can be of much greater help to the biologists. Furthermore this tool is supported with external information from different data sources that can provide the end user with the details of the proteins and more knowledge from different external sources.

Chapter 4

Structure based Information and Integration for the Alignment of Biological Pathways

Discussing VINEdb in the previous chapter it is known the information pertaining to the genes, proteins, diseases, pathways etc. are inter-related to each other. These pathways may be connected to different diseases and can also share similar proteins in different organisms. Therefore a system that can significantly utilize the rudimentary biochemical knowledge of the proteins involved and can also decipher the evolutionary relationship of the proteins and pathways can be of much greater help to the biologists. And a method that can align the biological pathways or networks using integrated information system that uses different protein information will be helpful to find the homologous pathways especially disease related pathways. The principles that govern metabolism are similar in any living cell even though an array of chemical reactions occurs. And it is result of their common evolutionary origin and due to the constraints imposed by the laws of thermodynamics. Many of the specific or unique reactions are common to all organisms differences primarily will be due to the different sources of free energy that support them. Comparing proteins, several works and algorithms have been developed to find out their relationship. They are compared at the sequence and structure levels. At a higher level the protein-protein interaction is widely studied by different groups. They have also developed several algorithms, applications for the network, pathway comparison. These works focuses on the information pertaining to the sequence, enzyme level and other functional levels. But none of these works focus to compare pathways using the structural level information. Here we present SignAlign (Structure based information integration for the alignment of biological pathways). It uses the protein structural classification information from different data sources such as SCOP, CATH along with PROCOGNATE and QSCOP being other external sources. Moreover it also uses

** Parts of the work have been published in 2006, 2007, 2008.*

the EC number classification information for the pathway alignment. Furthermore SignAlign is fortified with a visualization of the alignment results using a relational colouring pattern. This gives the user clear information about the pattern among the pathways along with the information of the protein and the information about its classification hierarchy in the databases. By comparing pathways it will help to know their homology. This is will be especially helpful for the biologists who can use it to know the relationship among the disease pathways.

4.1 Introduction

“The use of the term “homology” implies that a given similarity is a result of common ancestry” (Ahouse, 2002).

It is evident that when proteins share significantly more similarity with respect to some features more than it is expected by chance, the most prudent explanation is that they are homologous and they have descended from a common ancestor and as well share some other similarities. Some of the most commonly compared features are protein’s sequence and structure although some other features can be conserved and can be compared (Sierk and Pearson, 2004). And it is essential for accuracy in function prediction and for comparative genomics the homologs, the sequence that share common ancestry in identified accurately. Identification of homologs are very integral to the annotation of the genes, gene fusion analysis, prediction of gene function by various methods, including phylogenetic clustering, genomic context and phylogenetic interference. Moreover pairwise homology detection is a highly essential component for clustering approaches to protein family classification. It is obvious that all the applications try to exploit the following properties of homologous sequences: the genes that share common ancestry (a) located in the homologous chromosomal regions it makes the genes very suitable markers for comparative genomics (b) they tend to have similar structure and function (Song et al. 2008).

The key idea of sequential alignment was introduced by Needleman and Wunsch as a matrix approach which states that the optimal alignment is the best way through the matrix given a scoring function. This led to the development of several methods for protein sequence and structure alignments resulting in achievements and our understanding of protein similarities, their evolutionary relationships, functionality etc. At the same time according to the number of cases reported in literature, which is

against and unusual from the sequential point of view, structurally equivalent parts have different connectivity in the sequences of compared proteins. The alignments from these comparisons cannot be represented as a diagonal path through the matrix. The more interesting and surprising results came from the analysis performed is that non-sequential alignments have been found in large quantities in structurally similar proteins or to we can say the there are many alignments between highly structurally similar proteins for which the alignment matrix is not diagonal (Abyzov and Ilyin, 2007).

During evolution the protein structures and their functions are more conserved than sequences and further being classified by their structural similarity. The solved structures are regularly stored in Protein Data Bank (PDB) at the same time they are studied by several groups and classified and maintain their databases. FSSP (Holm and Sander, 1997), CATH (Orengo et al. 1997) and SCOP (Murzin et al. 1995) are the widely used databases and have their own way to compare and classify proteins, the resulting classification schemes are nevertheless largely consistent with each other (Getz et al. 2004). When compared the similarity between two proteins indicates and suggests a relationship between the structure, function and the evolution of the two proteins involved from a common ancestor protein. When either of the proteins is well characterized (i.e. in terms of structure and function) and further if the connection with a novel sequence is also established then it is possible that all the hard-earned biological data can be transferred to the newly discovered protein. This transfer depends on the degree of certainty how similar the two sequences are, as it is possible that even for distant relationships it is likely that the two proteins have the same overall structure (at the fold level). Furthermore it is also possible to suggest that they can have the same function. These suggestions can help to design the new experiments on the novel protein. It is for the same reasons as sequence comparison structures are also compared i.e. to know the homologous proteins, discovery of motifs (structural motifs) and for classification. And comparing structures can reveal interesting relations that are not possible to identify using the sequences alone. (Eidhammer et. al. 2004).

The interacting pairs of proteins should co-evolve and also they should maintain functional and structural complementarity and as a consequence such a pair of protein families shows similarity between their phylogenetic trees. Though the tendency of

co-evolution has been known for various ligand - receptor pairs it has not been studied thoroughly and systematically to the possible extent. In order to maintain the functional complementarity the interacting protein pairs should co-evolve as mutations in one protein should be compensated by mutations on the other (Kim et al. 2004). Further it is reasonable to assume the 'new' protein will also have that function when a protein of unknown function interacts with one or more protein of known identical function. Although the correlation can often be weak there are evidences states that properties of interacting proteins, such as the evolutionary rate, expression level and regulatory elements are more similar than to expect them as a chance (Stumpf et al. 2007).

In order to develop a drug against a disease, it is essential to select a protein which is linked to the disease which in a way suggests that it would be useful therapeutically to affect its function or expression. Rich source of information are provided which are of high importance for identifying potential drug targets especially of genome sequences and protein expression patterns are being generated by new high-throughput data sources. It is possible to pinpoint with differential genomics and proteomics by comparing the healthy and diseased animals or humans which particular protein is missing or dysfunctional or being regulated improperly or it is only expressed in affected cells. Not just a single protein, with the information about protein-protein complexes it is possible to target but a specific protein-protein interaction. With the knowledge of prokaryotic and viral genomes it is possible to support the identification of drugs against disease which are infectious. Metabolic pathways unique to micro-organisms are of particular interest and so are the proteins that take part in them. It is less likely that the drug which affects such a target will interact with a human homologue, with side effects consequently. Further there is a possibility of broad-spectrum antibiotics with proteins which have similar sequences across bacterial clades. Also, there is a warning of potential redundant functions by gene duplications conversely along with the concomitant insensitivity of the target inactivation. Moreover the expected stability of a therapy against the development of strains that are resistant is indicated by the knowledge of the relative evolutionary speed of various proteins, which include the rates of horizontal gene transfer (Lesk AM, 2005). Structural biology in combination with computational tools is playing a vital role in systems biology. Moreover structural information can help to identify homologues

that are not very evident from sequence data. Furthermore, structures can also be used to assess their immense capacity to bind drug-like molecules (Bakheet and Doig, 2009). Recent works have also proved the possibility of calculating the enzymatic kinetic parameters from protein structures (Stein et al. 2008) that could to a great extent fill the gap of non-availability of kinetic data which is required for the modelling and simulation of biological pathways /networks and is missing or available only under normal experimental conditions.

4.1.1 Structure versus sequence

Nearly all proteins have structural similarities with other proteins and in many cases, share a common evolutionary origin (Murzin et al. 1995). To understand the evolutionary relationship of pathways it is not sufficient to consider only the sequence level information or enzyme information. Detection of a structural similarity between proteins can often provide clues, since this is often accompanied by a similarity in function. This is often true even for very weakly similar sequences. Two proteins can have sequence identities below 10% and still perform similar function (Shah et al. 2005). It is obvious that proteins hold both in their sequence and structure the information which relate to their evolutionary origin and also the function and mode of action. When comparing distantly related proteins the valuable information is, in proteins the structure is likely to be more conserved than the sequence if the function of the protein is to be maintained. Moreover, clustering the proteins based on both criteria into families and further observing the conservation of their residues/structural elements and also their positions can highlight those aspects that are very characteristic or unique of a family (Stebbing and Mizuguchi, 2004). Also the number of 3D structures for proteins is growing at a rapid pace. In some cases biologically important and interesting similarities can be revealed comparing the 3D structures that are not detected by comparing the sequences. In bioinformatics protein sequence and structure comparative analysis play a vital role or they are the foundation stone. It is possible to infer the similarities of the biological function of proteins when there is sequence and structural similarities among the proteins both having associated with the evolutionary origin. At the same time with structural comparisons it was possible to identify the super-families of distantly related proteins more than sequence comparison (Holm et al. 2008).

4.1.2 Structure versus Enzyme classification

Enzyme classification is independent of the protein structure and its reaction mechanism. As a result the enzymes that catalyze the same reaction from a variety of mechanisms are placed under the same classification, leading to the structurally and mechanistically different proteins considered being identical (Dobson and Doig, 2005). Whenever a new 3D structure of a protein is determined it is essential to know about the other proteins that have similar structures that allow inferring the function of the new protein if it is not known (Rocha et al. 2009). Proteins that are involved in complex metabolic pathways are encoded by many genes and the functions of these pathways result from the interplay of the involved enzymes. The metabolic pathway variant of the particular organisms under the study must also be investigated along with the investigation of single genes in isolation, in order to understand the phenotype of any living system. This will lead to certain interesting questions such as, within the group of organisms or more elaborately what are the commonalities and differences? Further, if these groups are divided into group of organisms that are of interest into subsets, 'what are the features that are shared by one set of organisms while these functions are completely lacking in the other set' (Oehm et al. 2008).

Databases like KEGG, GO, UniPROT, DIP, BIND, SCOP, CATH etc. have lots of information about proteins, pathways, ontologies and more. But these databases give information about an individual protein, pathway or set of proteins or pathways in particular. The question is how we can utilize the valuable information to broaden our knowledge to compare the proteins and biological pathways and further know about their evolutionary and inter relationships. And it is well known that current and significant research addresses more on topics like prediction of pathways, network alignment and comparison of protein interaction networks across species and also about integration of databases.

4.2 Related works - alignment a common theme

A protein or set of proteins could be studied by the alignment of their sequences or structures. The sequence alignment could involve bringing in close the residues which can be similar or identical from two or more proteins. This could involve introducing gaps for the better alignment. And the structural arrangement involves the study of protein sequences and also by superimposing two or more proteins sharing same

structural property. All this are performed to find out the evolutionary relationship among proteins and to find whether they are homologous or not.

Alignment is a way of assigning correspondences that can preserve the order of the residues within the sequences. For this gaps can also be introduced. For e.g. given two strings of text:

One : *a b c d e*

Two : *a c d e f*

In this case a reasonable alignment would be

a b c d e – OR *a – c d e f*

In order to get the best alignment the criterion or criteria must be well defined. The above example defined the pairwise alignment. It is also possible to find large families of similar sequences and it is done by identifying the homologues in different species. Hence a mutual alignment of two or more sequences is termed multiple sequence alignment. Compared to the pairwise alignments multiple sequence alignment is much more informative and they can reveal patterns of conservation (Lesk AM, 2005). A biological pathway comprises of lot of enzymes and its products along with its substrates. Moreover they also interact with each other. Hence it is quite complicated and comparative analysis of pathways can be hardly done without a software system that is not well designed. Number of works has been done in the past for the comparative analysis of pathways like SEED, KEGG, Path-A, KAAS etc. No system is better than the other since the pathway analyses are complicated and also involve lots of tools and databases. Hence the user should choose a system of their interest that satisfies and fits into their area of research (Choi and Kim, 2008).

It is possible to identify certain common patterns of interactions (templates) among the similar pathways within a given set of pathways that are made of similar compounds yet behaving in a dissimilar manner i.e. functionality dependant, by some means of abstraction. This is based on the hypothesis that the pathways might have reached their present state after diverging from their primitive template and the divergence being dependant on the task (Reddy et al. 1993). Dandekar and his co-workers in 1999 proposed a pathway alignment comparative analysis method (**Figure 25**). This was applicable for the alignment of metabolic pathways in various genomes. This method and analysis yielded very crucial information about the evolution of

metabolic pathways and about their pharmacological targets and their applications in biotechnology. Using a combination of three different ways of comparison they studied the biochemical pathways

- (a) comparison and analysis of biochemical data
- (b) based on the concept of elementary modes pathways are analyzed and
- (c) comparative analysis of the sequenced genomes

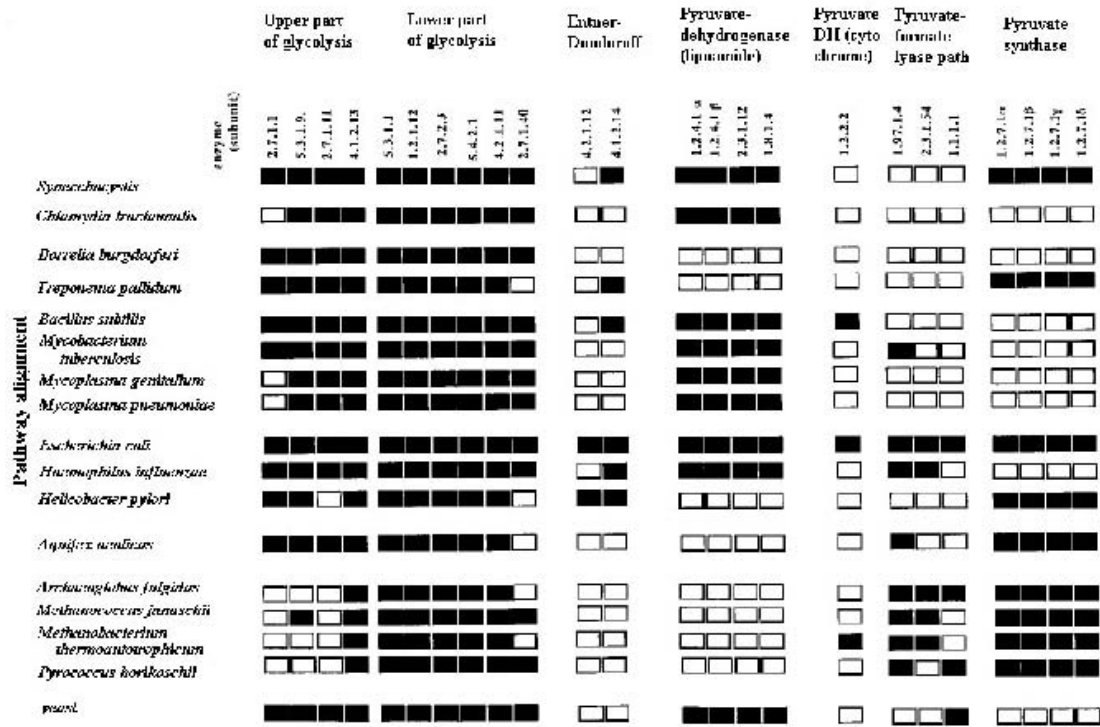


Figure 25: Comparative alignment method proposed by Dandekar et al., 1999

Later tools like PathAligner (Chen and Hofestädt, 2004.) and MetaPathwayHunter (Pinter et al. 2005) detect the conserved metabolic pathways using E.C. numbers. PathBlast (Kelley et al. 2004) searches for high-scoring alignments between pairs of protein interaction paths, for which proteins of the first path are paired with putative orthologs occurring in the same order in the second path. Netalign (Liang et al. 2006) used for comparative analysis of protein interaction networks. It compares a query protein interaction network with the targets combining interaction topology and sequence similarity thereby identifying conserved network substructures. The works of Hirsh et al. 2007 compare conserved protein complexes. ComPath is an interactive workbench for reconstruction of pathways along with annotation and analysis. Using an intuitive and interactive spreadsheet-style interface it allows the experts to perform

various sequence, domain and context analysis. It uses the spreadsheet style web interface to perform various analyses based on sequence information for enzyme and pathways comparison. Also, it helps to search the genomes for the prediction of the enzymes *de novo*. In comparison with the genomes of choice it also helps to annotate the genome. FASTA, Whole-HMM, CSR-HMM are the computational methods and PDB domain search are integrated into ComPath, to fill in the pathway holes or make *de novo* enzyme predictions. KEGG is the primary database suite integrated into ComPath, for pathways, sequences, functional classification and for compounds/reaction information. Motif and structural information are integrated from Pfam, Prosite, SCOP, SCOPEC, Superfamily and PDB databases in addition to sequence, structure and domain information databases (Choi and Kim, 2008). FMM (From Metabolite to Metabolite) is a web server which can reconstruct the metabolic pathways from one metabolite to the other among different species. It is mainly based on KEGG, UniProt and dPTM (Chou et al. 2009). A comparative table (**Table 7**) shows the different features of the various works and tool below.

Table 7: A comparative table showing the different features of the tools developed for alignment and analysis of pathways


	Dandekar et al.	Metapathway Hunter	PathBlast	NetAlign	PathAligner
Institute	EMBL, Heidelberg	Israel Institute of Technology	White Head Medical Institute	University of Science and Technology of China	Bielefeld University
Objective	comparative analysis of metabolic pathways	pathway alignment	network alignment and search tool for comparing protein interaction networks	comparative analysis of protein interaction networks (PINs)	alignment and prediction of biochemical pathways
Architecture	not known	software infrastructure	web application	web application	web application
Complexity	not available online	needs time for installation	only web browser	only web browser	only web browser
Web Interface	not available online	downloadable	complete	available	complete
Platform independent	not available online	yes	yes	yes	yes
License	free	free	free	free	free

All the earlier works focused on pathway comparison mostly using the protein sequence and functional level information or the EC number based classification or they are based on the protein interaction networks or protein complexes. It is understood from these and earlier studies, just focusing on the sequence or enzyme information does not help much in find the relationship among the proteins and pathways. Hence a new system is needed that can use some other information apart from this can be useful to compare biological pathways.

4.3 SignAlign

Knowledge about the three-dimensional structure reveals the details about the binding, catalysis and signalling (George et al. 2004). And “Knowledge of protein structure based either on prediction or on crystal structures provides a detailed level of information about protein function that is not available from other experimental data” (Teichmann et al. 2006). As a result of proteomics rapid development is expected in the coming decade and PDB (Protein Data Bank) is also growing fast. Hence it becomes evident to compare protein structures. There is no absolute consensus available about the best one even though good protein aligners are available to researchers. Result, the lack of knowledge of an objective way to compare protein structures (Rocha et al. 2009). It is usual that protein structural comparison methods are mainly focused on detailed structural alignment between two proteins. The quality of their alignment is gauged by the Root Mean Square Deviation (RMSD) which is the result of their optimized superimposition. It is believed SCOP maintains highly accurate classification results and proteins with structural relationships are grouped hierarchically at the fold level. It is more labour intensive, even though manual inspection is very accurate (Chi et al. 2009).

Most proteins contain more than one domain, multidomain proteins and this has been shown in the computational analyses of sequenced genomes. Due to processes of extensive duplication and shuffling domains the proteins are evolved. In proteins the generation of novel and complex functions is from a limited set of domain families, this has been facilitated by the modular nature of domains. Number of example are known or available that show the contribution of the individual domain function to the overall function of the protein or sometime in order to link two biochemical processes which are separated, the domain with specific function recombine. In the studies of



Bioinformatics Department
Bielefeld University

SIGNALIGN

Alignment and Prediction of Biochemical Pathways

- [Home](#)
- [Search](#)
- [Releases](#)
- [Help](#)
- [Contact](#)
- [Publications](#)

ALIGNMENT

The underlying concept of alignment and prediction of biochemical pathways applies a similarity measure on the basis of structural classification of the involved proteins. For that purpose a pathway is defined as a list of interacting proteins or the arrangement of the PDB ids representing a protein in a pathway occurring in a similar pattern.

Please insert your input data in the following manner:

```

pathway1 :: 1kee, 1h1s, 1j3kd, 1oa3, 1rjzB, 1vna
pathway2 :: 1c30, 3htsB, 1we6, 1cgs9, 1fh3
pathway3 :: 1t36, 1tro, 1b5e, 1vie, 1g6vk
    
```

At least enter two pathways for the alignment. Separate each pathway by line-breaks. Proteins must be represented by their PDB ids and in addition the structurally important part of the protein can be specified by the chain (i.e. 3htsB). For classification the protein must be available in SCOP or CATH database, respectively. If there are several classifications for a protein or chain then the alignment algorithm tries to choose the best one.

For more information and classification details please visit the [SCOP page](#) or the [CATH page](#).

Insert your pathways for alignment

```

pathway1 :: 1kee, 1h1s, 1j3kd, 1oa3, 1rjzB, 1vna
pathway2 :: 1c30, 3htsB, 1we6, 1cgs9, 1fh3
pathway3 :: 1t36, 1tro, 1b5e, 1vie, 1g6vk
    
```

Scoring method: SCOP PDB SCOP CATH EC

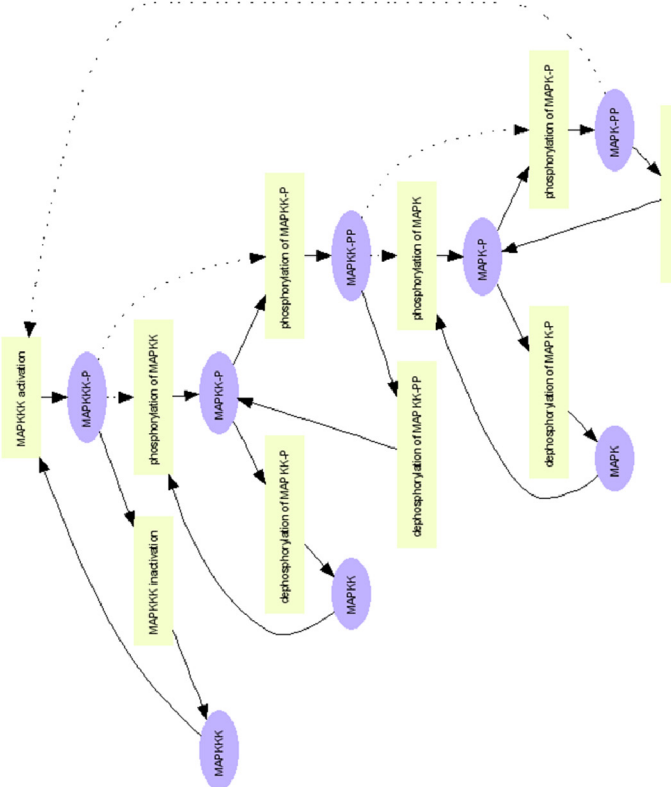
Image label: Small Middle Large

Gap Penalty:

Also the data can be inserted in the format of SBML. Please upload your SBML file after selecting it from your computer. Now, choose the proteins from the graphical representation of the pathways.

Load a pathway from a SBML file

Example:



<http://agbiotech.hak.uni-bielefeld.de> - Mozilla Firefox

Please select the corresponding PDB id from the table.

Protein name	PDB id	UniProt id	Entry name
mapk 2	4erk	p63086	mk01_rat
mapk 2	1tyo	p28482	mk01_human
mapk 2	3erk	p63085	mk01_mouse
mapk 2	3erk	p63086	mk01_rat
mapk 2	1gol	p63086	mk01_rat
mapk 2	1wzy	p28482	mk01_human
mapk 2	4erk	p63085	mk01_mouse
mapk 2	1pme	p28482	mk01_human
mapk 2	4erk	p63085	mk01_mouse

Figure 26: SignAlign showing the web based form where the user could interact with the system by entering the relevant PDB ids for the proteins involved in the pathways. Also the SBML file can be used for choosing the relevant PDB ids.

protein structure and function the classification of protein evolutionary relationship is important. Often it may be that even in protein that share little sequence similarity the structural similarities are present, as the structure conservation is stronger than the sequence conservation. Therefore in order to identify the distant relationship between proteins structure based classification is more effective than are sequence based methods and approaches (Han et al. 2007). Here a new approach to align biological pathways is proposed which is based on the rudimentary biochemical knowledge. A web-based tool SignAlign (**Figure 26**) is presented which can align pathways using information from an integrated database pertaining to the protein structural information along with enzyme and other significant information (Hariharaputran 2007, 2008).

4.3.1 Concept

The key concept of our approach is to use structural information of the proteins involved in the pathway as they are reliable compared to sequence level classification of proteins providing us knowledge about the homology of pathways. And it is imperative the structural information can throw more light in deciphering the evolutionary relationship among pathways and it was not possible with the previous works. Hence the need for a system that can utilize the protein structural information and the thought was born. At the same time it is to be taken care the data and the source is highly reliable which can help in the study and in knowing the protein structural information. Hence we choose to integrate data for the integrated database which comprises of information from SCOP, CATH, UniProt, Gene Ontology (GO), Protein Data Bank and enzyme information. Along with the above data it also integrates the external information from QSCOP and ROCOGNATE databases. All these data sources are widely used and have been appreciated by other works which focuses on structural aspects of the proteins. And SCOP and CATH have been in use for more than 10 years. Following is a brief about the different data sources:

SCOP – Structural Classification of Proteins

“Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification”
<http://scop.mrc-lmb.cam.ac.uk/scop/>

Structural Classification of Proteins database is a comprehensive ordering and complete description of the structural and evolutionary relationships of the structurally known and available proteins in PDB (Protein Data Bank). It follows a four level classification hierarchy system i.e. Class, Fold, Superfamily and Family. The domain is the fundamental classification in the experimentally determined protein structure. Further these domains are grouped together at different levels according to their sequence, structural and functional relationships. In the hierarchy if we proceed from the bottom to the top, the SCOP comprises the following levels: the protein Species, Protein, Family, Superfamily, Folds and Classes and each level has been categorized based on the dis/similarity of sequences, functions and structural features (Murzin et al. 1995, Andreeva et al. 2008)

CATH – Class Architecture Topology and Homology

“The CATH database is a hierarchical domain classification of protein structures in the Protein Data Bank. Only crystal structures solved to resolution better than 4.0 angstroms are considered, together with NMR structures. All non-proteins, models, and structures with greater than 30% “C-alpha only” are excluded from CATH. This filtering of the PDB is performed using the SIFT protocol. Protein structures are classified using a combination of automated and manual procedures”.
(<http://www.cathdb.info/>)

Class, Architecture, Topology, Homology (CATH) is a database of hierarchical classification **Figure 27** of protein domain structures that uses a combination of automated and manual approaches involving empirical methods, computational techniques, scientific literature reviews along with expert analysis. In CATH the domains are classified by the curators manually and are guided by the prediction

algorithms (such as structure comparison). Before being classified into homologous superfamilies according to both structure and function each protein structure is decomposed into one or more chains which are further split into one or more domains. In CATH each level i.e. Class, Architecture, Topology, Homology consists of domain classified based on their sequence, structural and functional similarities. The CATHEDRAL structure comparison algorithm is used which compares the structures and used to characterize structural diversity in CATH superfamilies and structural overlaps between superfamilies (Orengo et al. 1997, Cuff et al. 2009).

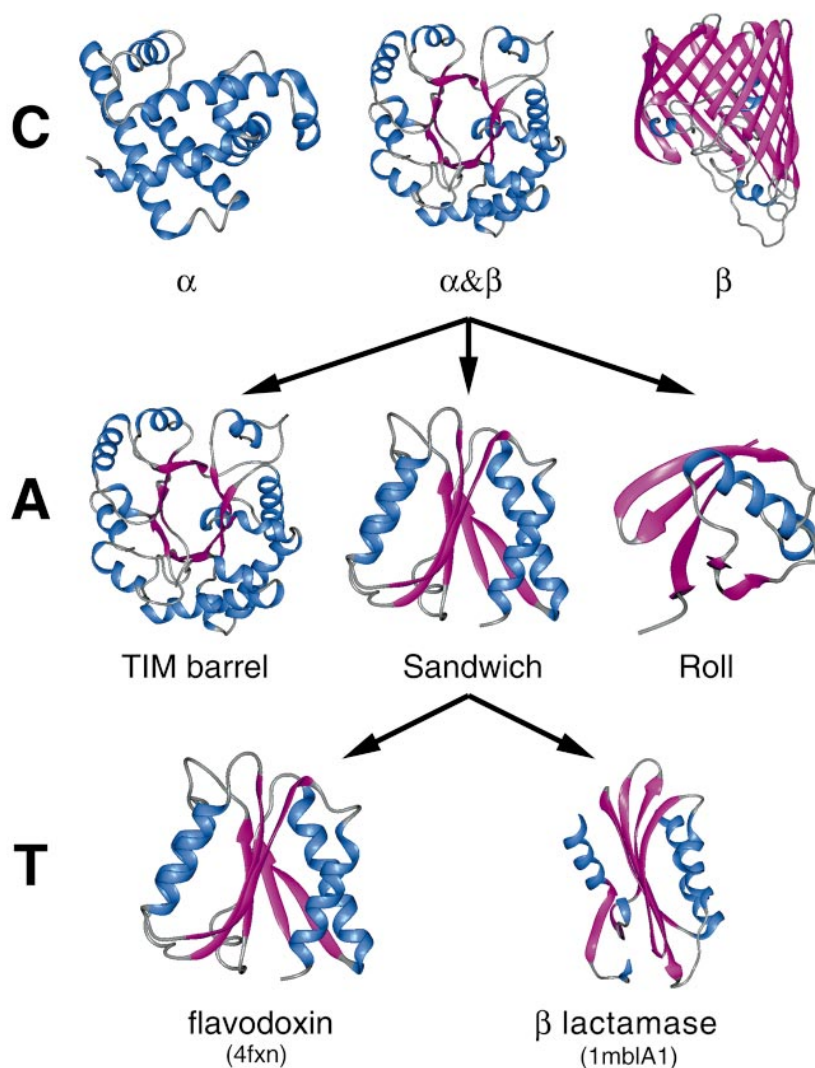


Figure 27: Schematic representation of the class (C), architecture (A) and topology (T) level in the CATH database. (Source: Orengo et al. 1997)

PDB – Protein Data Bank

Protein Data Bank (<http://www.rcsb.org/pdb/>) established at Brookhaven National laboratories (BNL) in 1971 is the universal repository containing the information about the 3D structures of proteins and nucleic acids and single world wide archive of structural data of biological macromolecules. These large molecules are found in all organisms and understanding their shapes help to understand how they work. Furthermore this helps to decipher the role of the structures of these molecules in human health and disease and also in drug development. At the beginning the Protein Data Bank had seven structures and every year there is increase in the number of structures. Currently it stores more than 55,000 entries. The rapid increase is due to the result of improved technologies in crystallographic processes in addition to the structures determined by nuclear magnetic resonance (NMR) and also due to the changes in the community views about sharing of data (Berman et al. 2000).

QSCOP

QSCOP :: Protein-Browser

Protein-Browser

Description
The Protein-Browser displays all chains and domains of a specified PDB file and how they are classified in SCOP, along with the granularity of the respective SCOP Families.

Usage
Search by PDB code in the search field below. The result page consists of two tables: One table lists all available SCOP domains for the given PDB code and the other table shows the granularity of the corresponding SCOP families.

1t36 <= PDB code (e.g. 1igi).

SCOP Domains of 1t36

Method: X-RAY, Resolution: 2.10, Release: 2004-09-21

Domain	Length	Fragments	SCCS
d1t36a1	153	1t36,A(403:555),0	a.92.1.1
d1t36c1	153	1t36,C(403:555),0	a.92.1.1
d1t36e1	153	1t36,E(403:555),0	a.92.1.1
d1t36g1	153	1t36,G(403:555),0	a.92.1.1
d1t36b2	228	1t36,B(153:380),0	c.23.16.1
d1t36d2	228	1t36,D(153:380),0	c.23.16.1
d1t36f2	228	1t36,F(153:380),0	c.23.16.1
d1t36h2	228	1t36,H(153:380),0	c.23.16.1

Figure 28: A detailed QSCOP information for 1t36 along with SCOP ids

QSCOP (<http://qscop.services.came.sbg.ac.at/>) follows a way to extend the SCOP by four more layers called Distant, Related, Similar and Equivalent (**Figure 28**). All these layers are defined based on quantitative structural relationships among the individual SCOP families. Sequence similarity and functional similarity though being criteria have been deliberately neglected. By following this approach QSCOP quantifies the structural diversity of SCOP families leaving untouched the integrity and contents of classic SCOP. Granularity otherwise called the structural diversity of classic SCOP families is being represented by the separate groups in the QSCOP sub-layers. As such QSCOP quantifies and defines structural similarity when two SCOP domains shares a certain number of structurally equivalent residue pairs. These values are based on the information obtained by the superimposition of two SCOP domains using different tools like ProSup and Sippl. Based on the percentage similarity of the equivalent residues to the larger domains the similarity of the pairs of structures is further expressed (Suhler et al. 2007).

PROCOGNATE

PROCOGNATE (<http://www.ebi.ac.uk/thornton-srv/databases/procognate/index.html>) is a database of cognate ligands dedicated for the domains of enzyme structures in CATH, SCOP and Pfam. The database consists of PDB ligands to the domains of the structures based on the CATH, SCOP (**Figure 29**) and Pfam databases classification. Ligands for this database have been identified using data from the ENZYME and KEGG databases and are further compared to the PDB ligand using the graph matching approach for the finding the chemical similarity. Further ligands are assigned to the enzymes structures which have EC numbers and also have known reactions in ENZYME and as well in KEGG (Bashton et al. 2006, 2008).

GENE ONTOLOGY

It has been made clear by genomic sequencing that all eukaryotes share a large fraction of the genes specifying the core biological functions. The knowledge of gene and protein roles in cells is accumulating everyday and they also change. The Gene Ontology project is a combined effort for dealing with consistent descriptions of the gene products in different databases. The consortium maintains a relational vocabulary in a hierarchical manner and describes the ontology of the gene products on three well organized levels, Biological processes “*refers to a biological objective*



PROCOGNATE

[Main Page](#) | [Help](#) | [Stats](#) | [Download](#)

1t36

Structure Title: Crystal structure of e. coli carbamoyl phosphate synthetase small subunit mutant c248d complexed with uridine 5'-monophosphate

Structure Header: Ligase

Associated ECs: 6.3.5.5 [S]

Chains:

Code Molecule

1. C Carbamoyl-phosphate synthase large chain, current chain [S]
2. D Carbamoyl-phosphate synthase small chain

Change domain classification: [CATH](#) | [SCOP](#) | [Pfam](#)

Domain	Superfamily	Superfamily Name	Ligand bound by > 1 domain	PDB ligand (code residue chain, name)	Cognate Ligand	Similarity
1.	c.30.1 [S]	PreATP-grasp domain	Y	K 24 -, Potassium ion [C][S]		
			N	NET 30 -, Tetraethylammonium ion [C][S]		
			N	CL 34 -, Chloride ion [C][S]		
2.	d.142.1 [S]	Glutathione synthetase ATP-binding domain-like	N	ADP 20 -, Adenosine-5'-diphosphate [C][S]	ADP [R][S][L]	1
			N	MN 21 -, Manganese (II) ion [C][S]		
			N	MN 22 -, Manganese (II) ion [C][S]		
			N	K 23 -, Potassium ion [C][S]		
			Y	K 24 -, Potassium ion [C][S]		
			N	PO4 25 -, Phosphate ion [C][S]	Orthophosphate [R][S][L]	1
			N	CL 32 -, Chloride ion [C][S]		

Figure 29: PROCOGNATE information for 1t36 using SCOP based classification to which the gene or gene product contributes”, Cellular components “refers to the place in the cell where a gene product is active” Molecular functions “is defined as the biochemical activity of a gene product” in a species in an independent manner.

The keyword databases of many gene and protein are linked to each node in the GO ontologies along with other kind of information. The aim of the project is to create a dynamic and controlled vocabulary that can be applied to all the eukaryotes even when the roles of the gene and protein is changing and still accumulating (The Gene Ontology Consortium, 2000).

In SignAlign for the alignment of biological pathways the scoring method implemented is based on the classification and amount of information content shared between the items arising from the same parent node. Further this information is incorporated as a factor to the alignment scoring method along with a gap penalty, when needed. The multiple alignment algorithm is based on a progressive alignment algorithm method. It works by constructing successive pairwise alignment and in each step the number of sequence that is aligned is reduced by one. And in the case of pairwise alignment the Needleman-Wunsch algorithm is modified and similarity score is produced by comparing the proteins based on SCOP, CATH or EC number classification.

For the efficient retrieval of the optimal alignment of two sequences the algorithm uses a dynamic programming approach. The dynamic programming matrix F is constructed and it used to get an optimal alignment in the following ways

- (a) F is an n - m -matrix with n the length of sequence 1 and m the length of sequence 2, respectively.
- (b) $F(0,0) = 0$ and the first row and column are initialized as follows: $F(i,0) = i * -d$ and $F(0,j) = j * -d$, where d is the gap-penalty
- (c) the whole matrix F is filled with the recurrence:

(where $s(p_1, p_2)$ is the score between protein p_1 and p_2 , respectively)

- (d) the value of the final cell $F(n,m)$ is by definition the optimal score of the alignment
- (e) by trace back from the final cell to $F(0,0)$ one can retrieve an alignment with the optimal score

Progressive multiple alignment of N pathway sequences is then carried on by the following algorithm:

- (a) from the set of pathways an all-against-all pairwise alignment is done and the alignment with the best (highest) score is chosen
- (b) a pattern is built from the chosen alignment
- (c) both sequences of the alignment are removed from the set of pathways and further found pattern sequence is added to the pathway set
- (d) go to 1 until all pathways are included in the alignment

This is a heuristic algorithm and hence it is not guaranteed to find the optimal multiple alignment and its runtime is polynomial.

4.3.1 Architecture

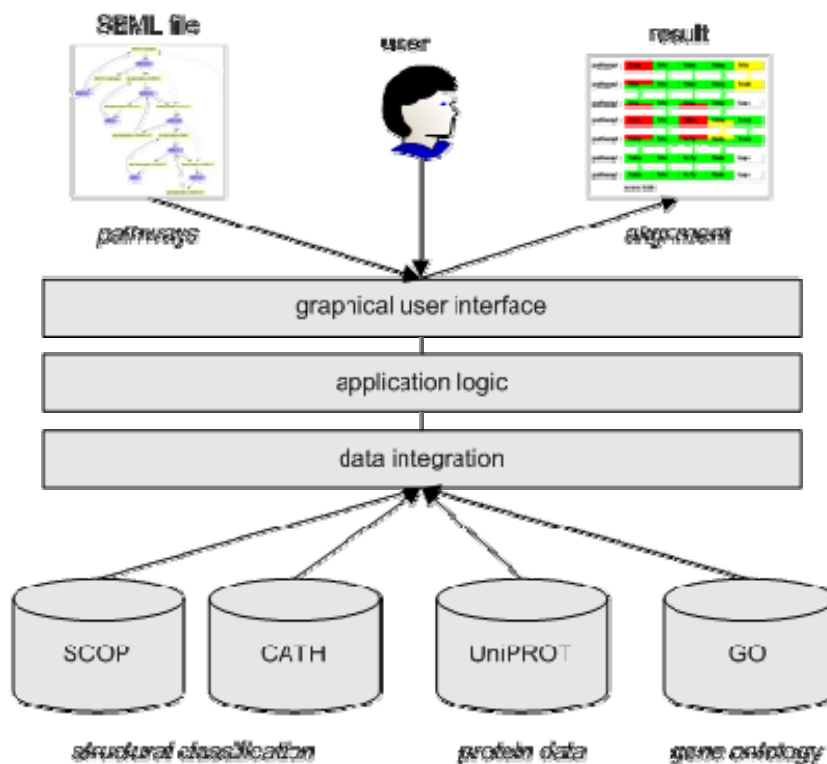


Figure 30: A schematic representation of the 3-tier system architecture in SignAlign

SignAlign has a three tier architecture consisting of the data integration layer, application layer and a graphical user interface (**Figure 30**). The data integration layer at the base of the system integrates suitable information from SCOP, CATH, UniPROT and GO. Also the enzyme (EC number) information and Protein Data Bank (PDB) are integrated into the system that is described in the previous section. Based on the integrated data an application logic layer performs the alignment algorithm and controls integration of data as well as user interactions. By dint of the graphical user

interface scientists can enter their pathways for alignment. Furthermore the user will be also be able to select the pathway elements from SBML files available from different research groups or their own SBML which are compatible with SignAlign and that are stored in their local computers. The resulting alignment is visualized by an image and it is based on the algorithm specific to the method i.e. SCOP, CATH or EC number based classification using a relational colouring pattern. This is further supported with more external information.

4.3.2 Implementation

The underlying concept of alignment and prediction of linear pathways applies similarity measures on the basis of structural classification of involved proteins. In this work a pathway is defined as the set or ordered list of interacting proteins sharing similar structural and domain features with their counterparts. An alignment algorithm retrieves the corresponding structural information of the proteins from the database accompanied with appropriate visualization for clarity. Having this as the base, the difference between any two pathways can be quantified as structural similarities or identities between closely and distantly related proteins, by comparing the position of the protein domains in the classification hierarchy.

Having been implemented in Java makes SignAlign platform independent. And the web-based graphical application is implemented with JavaServerPages and runs on Apache Tomcat web server. The integrated database is supported with MySQL which stores the information from SCOP, CATH, UniProt and Go along with EC number and PDB information. The graphical images are generated using GraphViz (Gansner et al. 1999) the graph visualization software. This helps to create the dynamic images at run time for the alignment algorithm. Through the web-based graphical interface it is possible to interact with the system and allows the end user the access to the integrated information. The web pages of the system can be accessed using any common web browser. In addition the tool provides a search engine that mine the related inbuilt information about the proteins and about their classification scheme both in CATH and SCOP along with the relevant UniProt and PDB information. An SBML converter facility allows the import of existing SBML files directly into the system. Users can select the relevant PDB structure of the protein involved in the pathways from the projected list.

Different systems discussed earlier (**Table 7**) use the diverse information and generate their results in different ways. Having considered the various properties SignAlign is developed in-house for the alignment of biological pathways and it is available to the community for free. All these systems have been focussing on the alignment of metabolic pathways but they did not use protein structural information. SignAlign is able to align biological pathways using protein structural classification information. The comparative **Table 8** shows the various properties of different works along with SignAlign.

Table 8: A comparative table showing the different features of the tools developed for alignment and analysis of pathways along with SignAlign

	Dandekar et al.	Metapathway Hunter	PathBlas t	NetAlign	PathAligner	SignAlgin
Institute	EMBL, Heidelberg	Israel Institute of Technology	White Head Medical Institute	University of Science and Technology of China	Bielefeld University	Bielefeld University
Objective	comparative analysis of metabolic pathways	pathway alignment	network alignment and search tool for comparing protein interaction networks	comparative analysis of protein interaction networks (PINs)	alignment and prediction of biochemical pathways	alignment of biological Pathways
Architecture	not known	software infrastructure	web application	web application	web application	web application
Complexity	not available online	needs time for installation	only web browser	only web browser	only web browser	only web browser
Web Interface	not available online	downloadable	complete	available	complete	complete
Platform independent	not available online	yes	yes	yes	yes	yes
License	free	free	free	free	free	free

4.3.3 Application

The user can interact very easily using the web interface shown in **Figure 26**. Using the form can bring in the data i.e. proteins represented as PDB ids into the form as an array or chain in order to represent the pathway. Each PDB id must be separated by a

comma. The example of multiple pathway provide in the web page gives more information about the input. Also in the home page details are provided about the different scoring methods i.e. SCOP, CATH and EC number that are implemented in the system along with the gap penalty. Further the users can also adjust the size of the image that will be generated as result for the alignment and they will be able to set the required label. Along with these there is a facility provided to upload SBML files. At the side panel shown is the search engine which itself has a features will give the facility to mine the information of the protein.

The input for SignAlign is a set of proteins. The user can represent the pathways of their interest as a set of proteins represented by the PDB ids and type in into the interactive form (**Figure 26**) presented by the system in the following manner

Pathway 1:: 1kee, 1h1s, 1j3kD, 1oaj, 1rjzB, 1vna

Pathway 2:: 1c30, 3htsB, 1me6, 1cg9, 1fh3

Pathway 3:: 1t36, 1tro, 1b5e, 1vie, 1g6vK

It is also possible to enter the PDB along with the chain id if the user already know or can be searched using the search engine provided in the system. The user will have to have the knowledge about the pathways and about the proteins and about the PDB structures in order to feed into the interactive form. It is imperative some of these proteins are also enzymes. Following table shows some of the EC number, proteins (PDB ids) and associated pathways. And some of the proteins and pathways are further associated with cardiovascular and neurodegenerative diseases, apoptosis etc. in a direct or indirect manner. Hence the above set of proteins is chosen in order to represent a pathway considering them as direct protein-protein interaction.

EC number	PDB ids	Pathways
6.3.5.5	1c30, 1kee, 1t36	Pyridine metabolism, Glutamate metabolism
1.5.1.3	1j3k, 1b5e, 1vie	Folate biosynthesis
4.2.1.1	1g6v	Nitrogen metabolism
3.4.23.39	1me6	--

The algorithm requires minimum two pathways to compare and will be able to perform only linear alignment of pathways. More than a pair of pathways is always accepted by the system for the multiple alignment of biological pathways. Moreover, SignAlign accepts SBML (www.sbml.org) files as input and the end user can insert the PDB ids through the SBML file directly into the system. The built in facility within the system converts the SBML file into interactive file thereby allowing the end user to click on the participating proteins that is showing in the map. This is followed by fetching the relevant PDB ids of the involved proteins from the underlying database and the list of the PDB ids is shown to the user. The user now has the discretion to select the appropriate PDB ids and insert them into the interactive form of SignAlign. The search engine in SignAlign can provide more information pertaining to the proteins like classification information from SCOP, CATH and EC number by providing PDB id of the proteins and name. This will be more helpful to the user while deciding the input, the nature of the pathways and relevant PDB ids. Based on the input or set of pathways feed by the user, the algorithm performs the alignment. The output is an alignment of a given pair / set of proteins and pathways by the end user. The result is based on the similarity of structural or enzyme information that is shared between the pathways and based on the scoring method, i.e. SCOP, CATH and EC number. From a hierarchical point of view it will be a top-down approach in order to find the homologous pathways. Further the result of SignAlign alignment is fortified with a relational visualization (**Figure 31**) approach embedded with a colouring pattern. The colours are associated with the colours given in the legend of the tool which gives the information of the classification scheme followed by the chosen scoring method. Apart from this the gaps are indicated by colour white, missing proteins are indicated by black and mismatch by colour red. The alignment result is interactive and can give the end user the information about the protein involved in the alignment from the Protein Data Bank (www.rcsb.org). In the example shown the highly conserved proteins in the homologous pathways that satisfied all the four levels of hierarchy in the chosen method are shown in green. The tool is further supported with an option to choose the Gene Ontology (GO) terms and has links to the external information from QSCOP and PROCOGNATE associated with the proteins.

[Home](#)
[Search](#)
[Releases](#)
[Help](#)
[Contact](#)
[Publications](#)

VISUALIZATION

The results of the global alignment are generated based on the levels of similarity. Having this as the base, the differences between two pathways can be quantified as structural similarities between closely and distantly related proteins by comparing the position of the protein domains in the hierarchy and inserting this information as a factor to the alignment scoring.

The scoring method for this alignment is based on the SCOP Classification String (SCCS). Proteins are colored according to the level of their similarity. Gaps are indicated white and red signifies a mismatch, invalid or not classified PDBs are given in black.

pathway3 : 1t36 1tro 1b5e 1vie 1g6vK ---

pathway1 : 1kee 1h1s 1j3kD 1oaj 1rjzB 1vna

pathway2 : 1c30 3htsB --- 1me6 1cg9 1fh3

Pattern : a.92.1.1 a d.117.1.1 b b.1.1 g.3.7.1

Generated by SignAlign

Gap: ---	Fold: a.1.-
Missing: 	Superfamily: a.1.-
Mismatch: -.-.-	Family: a.1.1
Class: a.-.-	Identity: identity

Options for the visualisation

Classifications:

SCOP CATH EC numbers

GO-terms:

Cellular component Biological process Molecular function

Links:

QSCOP PROCOGNATE-SCOP PROCOGNATE-CATH

Figure 31: The visualization method followed by SignAlign to display the alignment result based on the SCOP classification scheme. It is supported by other information and colour legend and its significance. Also shown is the option to choose the external information.

Figure 32: The related information table from different sources such as Protein Data Bank (PDB), SCOP, CATH, Gene Ontology and other external links are provided as table to the user. The table follows the similar alignment colour pattern.

A table (**Figure 32**) will be displayed with the information and it follows the similar colouring pattern of the alignment result which is generated by SignAlign. The relevant information available from these external sources pertaining to those proteins are given in the table like the classified ontology terms from Gene Ontology and E.C. numbers associated with them.

4.4 Limitations

- The algorithm will be able to perform only linear alignment of biological pathways.
- The system depends on the external database information for the comparison of pathways.
- It is not possible to provide the PDB structures for all the proteins in the pathways since their structures are yet to be solved.
- The PDB structures are not available for all the proteins hence there is a difference in the number of available PDB structures and known proteins.
- The external sources (SCOP, CATH, EC number) used for protein comparison has their own limitations.
- There is also discrepancy in the numbers of structures and classification information available in PDB versus SCOP, PDB versus CATH, PDB versus EC number and SCOP versus CATH.

4.5 Summary

In a group of pathways they may have similar compounds but still behave in a dissimilar way especially due to their functionality. Yet it is possible to identify some common patterns as they would have diverged from a common or primitive template. The evolutionary relationship between the pathways can be analyzed at different levels of protein information i.e. sequence, structure, function and also using enzyme based methods. Most of the earlier works have been focusing on the sequence level information for the multiple alignments of biological pathways (result of protein-protein interaction). Also, enzyme based methods have been proposed to decipher the evolutionary relationship among pathways. Our approach enables structure based

alignment of proteins and prediction of linear biochemical pathways to know their evolutionary relationship. In SignAlign, we utilize the well defined and analyzed structural information from SCOP and CATH which classifies the protein based on the hierarchy level approach. By this approach it is possible to relate the evolutionary relationship between each protein that participates in the pathways. Moreover the approach is further supported with enzyme classification information (E.C. number) which also substantiates the findings. Also, the inclusion of external information from Gene Ontology, PROCOGNATE and QSCOP further supports the concept of deciphering the evolutionary information of biological pathways using protein structural classification information. The work can throw more light into their protein relationship and pave way for future methods of pathway evolution based on the protein structural information. SignAlign, is a tool for the comparison of biological pathways using structural information. But it can be deemed as a futuristic approach since it is not possible to provide the PDB ids for all the proteins which participate in the biological pathway. And it may take some time to solve the structure of all the proteins which will be useful for the protein structure information based analysis of biological pathways.

SignAlign is available at <http://agbi.techfak.uni-bielefeld.de/signalalign>

4.6 Outlook

In SignAlign the protein structural information plays a key role. Hence it will be valuable to integrate more protein structure information from different information systems such as Pfam, FSSP, PDB etc into the integrated database. This could lead to new and valuable information getting integrated into the system. Also, it is possible to integrate new methods or algorithms for alignment that can decipher and give more information about the biological pathways and it will be helpful for finding disease related pathways thereby giving new insights about their homology. It is known that biochemical parameters/kinetic parameters are not available for constructing biological pathways and it is not easily available under normal conditions. Sometimes if it is not possible to generate the parameters by experimental methods it is possible to generate the kinetic data through other methods like using protein structural information. This gives a new dimension and it is possible to expand the knowledge of biological pathways using structural information for the modelling and simulation of biological pathways/networks.

It was obvious from the previous chapter that biological pathways can share significant information at different levels at the same time it is also not possible to think that biological pathways can act alone or they are not connected to different information. The information in turn is related to the rudimentary biochemical / molecular data and has significant role in various diseases. Also, this could lead to the complex networks which comprises of several proteins, genes, enzymes, relations, reactions etc. Thinking a biological network with these and more information can throw more light in the areas of pathway modelling, which could possibly imitate the pathway function and also to generate significant hypotheses. The forthcoming chapter will deal with the semi-automatic construction of biological networks using an integrated database approach that further allows large scale modelling and simulation of biological networks using different Petri net based software suites. Moreover it emphasizes that it is not always possible to construct / reconstruct biological pathways manually or by hand they being complex. Hence a tool which can construct pathways on the fly can be great asset and useful for biologists and can be useful for the CardioWorkBench and other similar projects.

Chapter 5

Petri net based reconstruction and visualization of biological pathways using integrated molecular data

Enormous works are done in the past by different research groups in order to model and understand the complex functioning of the biological networks. And it is well known in order to understand the functioning of these networks they have to be studied very closer. For this efforts taken by previous works have to consider the myriad of biochemical reactions that occurs leading to changes in the state. And they emphasized on the need for kinetic parameters which is not available in most of the cases. This led to the change in the way networks could be modelled and understood. Some of them moved towards qualitative methods leading to the construction of smaller pathways. At the same time some of the recent works have shown that it is possible to generate the kinetic parameters from protein structural information. Considering the previous works and also knowing some of the bottlenecks a tool to construct large scale biological network is proposed. The tool MoVisPP will be able to generate large scale networks which are not dependant on the biochemical parameters and will be able to construct networks based on KEGG pathways. Further it will be able to integrated more information which is fetched from the database which supports the tool. The salient feature of MoVisPP, it will be able to construct the Petri net based networks integrating diverse information. Further it allows to export the generated networks in different formats such as SBML, CellML etc and to import them to different software suites such as Cell Illustrator. By this method it will allow the user to interact with the network and also to expand and model large scale networks incorporating the biochemical parameters to the different entities involved.

* *Parts of the work have been published in 2007, 2008, 2009.*

5.1 Introduction

Cell is composed of various molecules such as DNA, mRNAs, non-coding RNAs, proteins, modified proteins, molecular complexes, etc. forming complex networks which are physically located in the cell and are wired by numerous biochemical reactions. Moreover, cells of different functions act in concert in an organism, and the organism has also several systems that answer to the stimuli and also to various other molecules present outside the cells and in the environment. To model and simulate the biological systems in the cell and body and further to investigate and simulate them on computer there have been several challenges. With the advancement of biological measurement technology the challenges in the systems biology are now being accelerated. The recent high throughput analyses are generating a vast amount of heterogeneous data for the information on “when”, “where”, “who”, “what”, “which” and “how” of biological systems. At the same time data that are not possible through high throughput experiments and are obtained by traditional methods, is being accumulated in the databases and published as literature. This trend in systems biology requires inevitably computational information systems and tools and to enhance our abilities for understanding life – as system (Nagasaki et al. 2005).

The actions of integrative biology try to understand how the biological system components interact and how their interactions give rise to emerging functions and their systematic level behaviour. To support this are numerous experiments that produce varied data sets at different levels of organization that includes the genome, proteome and the metabolomic data along with them are newer computational techniques and the creation of various innovative devices. A mathematical model is the centrepiece of these diverse techniques that utilizes the experimentally observed data and further formulates them as theoretical structures that can be interpreted and validated by computational means (Wu and Voit, 2009). The post-genomic era paved the way and brought very high opportunities that can bridge and bring together disparate disciplines in the natural sciences. It was possible to probe the cellular function at multiple levels due to the high-through put techniques and also with the wide availability of large datasets that comprises of annotated genomes to organism-level maps of protein interactions and cellular metabolism. The reductionist approach that stated the behaviour of a system can be understood and predicted only from the detailed knowledge of the elementary constituents of a system can be correlated

directly for the dramatic development of natural sciences in the last century (Almaas, 2007). In many disciplines such as theoretical physics and technology and sociology and in humanities, networks have taken a prominent position in the past decade. In biology, particularly in systems biology they have gained particular prominence and descriptions based on networks have a fundamental role especially with the attempts for combining system-wide biological information with predictive modelling (Stumpf et al. 2007).

It requires more input data than what the current technologies can offer for a complete system-level understanding of any biological process. We imagine intuitively that a system could be modelled best only after we know all the molecules that are involved, their concentrations, what will be the effect on their neighbours by each individual part, how they fit together and how over the time their dynamic parameters such as interactions, concentrations and mechanics change. Given the fact that for measuring many of these parameters are still not yet known or on the drawing board or do not know if they exist at all it seems highly unrealistic. The question arises, whether it is essential to know all the details of a process in order to have a predictive model or to develop a useful systems-wide understanding. It has been suggested from the analysis of protein interaction maps it is possible to derive the initial, rudimentary models of biological networks even when sparse data can be used (Uetz and Finley, 2005). Moreover it is made feasible to collect the large data sets on protein activity and abundance with the enormous availability of high-throughput and multiplex techniques which quantify signalling and cellular responses. For systems biology it is a paradox as these large data sets themselves bring less understanding and more confusion. For this reductionist approach make more sense simplifying things down to the point. At the same time systems biology is well equipped with an efficient tool that can handle complexity with ease: *computation*. In order to understand and improve without much complexity the computation models must project the underlying mechanism and reflect the experimental data. Modelling can be approached either based on prior understanding of the involved molecular mechanisms or without having to make any assumptions of the mechanisms that underlie models can be constructed solely analysing the data *per se*. Known as “data driven models”, often reveal new surprising and unanticipated biological insights and

allow the multivariate biological measurements to become tractable to our intuition (Janes and Yaffe, 2006).

The amount and availability of biological data is increasing everyday from molecular biological networks that not only promises a deeper understanding, alongside it also confronts researchers with the problem of combinatorial explosion. The amount of quantitative data, such as enzyme kinetics is lesser compared to the qualitative network which is growing much faster. Because of ethical reasons or due to the limitations of experimental methods it is impossible to measure the quantitative data in most cases. This led to the rise and availability of qualitative data such as interaction data which was not sufficiently utilised for modelling purposes until now. Each day there is development of new approaches and methods but the application of these methods is often restricted due to the complexity of data. And it is made possible to explore the static and dynamic qualities of the qualitative data by the biochemical Petri nets. Extreme pathway analysis, elementary mode analysis and Petri net invariant analysis are established methods for the qualitative analysis of biochemical networks. Giving insight into the basic system behaviour, a model can be checked using these qualitative approaches for biologically meaningful behaviour and for its consistency further allowing the validation of the model. The metabolic networks are analyzed using the first two methods whereas the Petri net theory is applied in addition for signal transduction and gene regulatory networks and their combination (Grafahrend-Belau et al. 2008). Petri nets are a discrete event simulation method that is being developed for the systematic representation, in particular for their concurrency and synchronization properties (Hardy and Robillard, 2004).

In order to get deeper insights into the functioning of the complex biological networks mathematical models are increasingly used. Among the various other methods employed for the modelling and analysis of molecular networks Petri nets have emerged as a promising tool (Chaouiya, 2007). And it is imperative that an integrated understanding of molecular and development biology must consider the molecular species that are involved in large number and as well as very low concentrations of many species *in vivo*. Being a mathematical formalism developed in computer science the stochastic Petri nets can express the molecular interaction networks as quantitative stochastic models (Goss and Peccoud, 1998). Petri nets are bipartite directed multi-graphs consisting of nodes of two types called places $P = \{p_1, \dots, p_n\}$ and transitions

$T = \{t_1, \dots, t_n\}$ and directed arcs that connect only nodes of different types and are weighted by natural numbers. Places are for modelling typically passive elements such as conditions, states or biological processes i.e. chemical compounds as proteins. Transitions represent active system elements events, chemical reactions as deactivation /activation. Places are represented as circles and transitions as rectangles in graphical representations. Transitions drawn as flat rectangles are called input transitions and are without preplaces. The casual relations between the active and passive elements are described by the arcs which are drawn as arrows and if it is larger than one, they are labelled with their weights. An event is connected by the arc along with its preconditions in order to trigger an event which must be fulfilled and when the event takes it will be fulfilled with its postconditions. With tokens residing in places the fulfilment of a condition is realised. Different degrees of fulfilment are indicated by any integer number of tokens that is carried by a place in a discrete net. A transition may fire, if all preplaces of a transition are sufficiently marked with tokens. According to the weights given to the corresponding arcs, when a transition fires, the tokens that are the dynamic elements of the system, are added to its all postplaces that are removed from its preplaces. From a biological standpoint, the very presence of tokens in places is an indication of the presence of chemical species and its concentration and also it indicates its threshold (concentration above a certain level) in the cell (Sackmann et al. 2006). Further is possible to make the biologists understand intuitively the features of biological pathway and its intrinsic structure using Petri nets and further helping them to model mechanistically and also to simulate larger biological networks. This is possible using Cell Illustrator. With Cell Illustrator (<http://www.cellillustrator.org/>) biologists can draw, model, elucidate and simulate the most complex systems and biological processes. Further, the software tool allows researchers to model metabolic pathway, signal transduction pathways, regulatory pathways and to model dynamic interactions of different biological entities such as DNA, mRNA and also proteins. Furthermore, it allows visualizing these pathways and also interpretation of the experimental data and hypotheses testing along with the provision to generate model diagrams and simulation result charts for publication purposes. Cell illustrator is a software tool for modelling and simulation of biological pathways based on the concept of Petri net combined with an XML format called Cell System Markup Language (CSML) describing biological pathways for simulation (Nagasaki et al. 2005).

5.2 Reconstruction, modelling and visualization of biological networks

Enormous amount of data are produced by the methods of biotechnology which are to be analyzed and stored and several database systems are available in World Wide Web for the genes and proteins. In the field of molecular biology this view point allows the investigation of metabolic processes. In order to understand the molecular logic of cells the metabolic processes involved must be analyzed both in qualitative and quantitative terms. Important methods in this case are modelling and simulation, that will influence at the microscopic level both medicine and genetics (human). Several electronic systems are available including the KEGG after the collection of the biosynthetic processes by Boehringer Company which represented the metabolic map in the graphical form. The question that arises now is how to include the dynamics in these processes (metabolic pathways) representations for the achievement of a more comprehensive, realistic analysis and biochemical reaction synthesis. It is not possible to understand and analyze mathematically the complex biochemical networks that involve cascades of enzyme reactions, but by complementing experimental studies with real biochemical systems it is possible to simulate them computationally and to determine how they work. For a more accurate representation of the real systems and to obtain the best agreement with the experimental data the parameters of such models can be modified. It is possible to study computationally the changes that happen in the concentration of the species of a reaction pathways which is not possible or may not accessible in the real systems. Such an environment can be deemed as “Virtual biochemical laboratory” which can be used to explore conditions which are not practical to test experimentally in the real world systems. The field of biotechnology modifies the phenotype by the recombination of the genotypes using theoretical tools and methods of molecular genetics. And the field also comprises the methods of informatics and medical informatics (Hofestädt and Scholz, 1998).

In systems biology the fundamental task is pathway reconstruction which leads to the ultimate goal of full-scale *in silico* simulations and there is lack of data for such reconstruction. Reverse engineering, means the inference of signalling, metabolic or gene regulatory pathways using the experimental data takes the centre stage in systems biology. To develop *in silico* models as concise representations of biological

systems experimental data would provide the sufficient detail. The behaviour of the real systems is recapitulated by the models which also serve as integrated tools in which essential components must be assembled and function together. With the models it is possible to predict and provide insight into how simple interactions give rise to complex behaviours. It is not essential that models must be fully detailed to achieve these goals. To a greater extent all models are abstractions. To test hypotheses and best direct limited experimental resources to collect the most salient data even simple and qualitative model play significant role. It is wise to make the best use of the data at hand to built complex models, beginning with more simple models. Large scale models developed earlier are certainly integrative in nature as the data from different sources are culled. From the topological or semi-quantitative methods such as Petri nets valuable insights are gained in some of the cases. At the same time rate or equilibrium constants are to be measured or they must be extracted from literature for more quantitative models (Rice and Stolovitzky, 2004).

With the advent of technologies generating huge amount of biological data it is not possible to focus on one single source or data for performing research, rather it require an interdisciplinary approach to understand and appreciate the knowledge and the source. Hence there is a need of a combinational approach that can enhance our knowledge as well as to bring in the information from different sources. But at the same time there are limitations as not at all the data sources provide enough detail pertaining to the protein, pathway kinetic and physico - chemical properties or parameters of the participating enzymes etc. So, there is a need of the source i.e. a data warehouse that can bring in the diverse information.

In order to understand and improve the complex biochemical processes computer aided possibilities are necessary. Moreover with the growth in the amount of biologically experimental possibilities and in the rapid explosion of related biological data, it has become essential to transmit the data in a very simple, analyzable and possibly validated models. In biochemistry the often used kinetic models are based on differential equations. Based on the experimental data and also had to prove and for the construction, analysis and simulation of a model, bio-scientists need theoretically well-founded and practicable methods. But these models require kinetic parameters in order to analyze a biochemical process or system. But the lack of reliable data or inconsistencies in the used kinetic model makes the results unreliable. This persuades

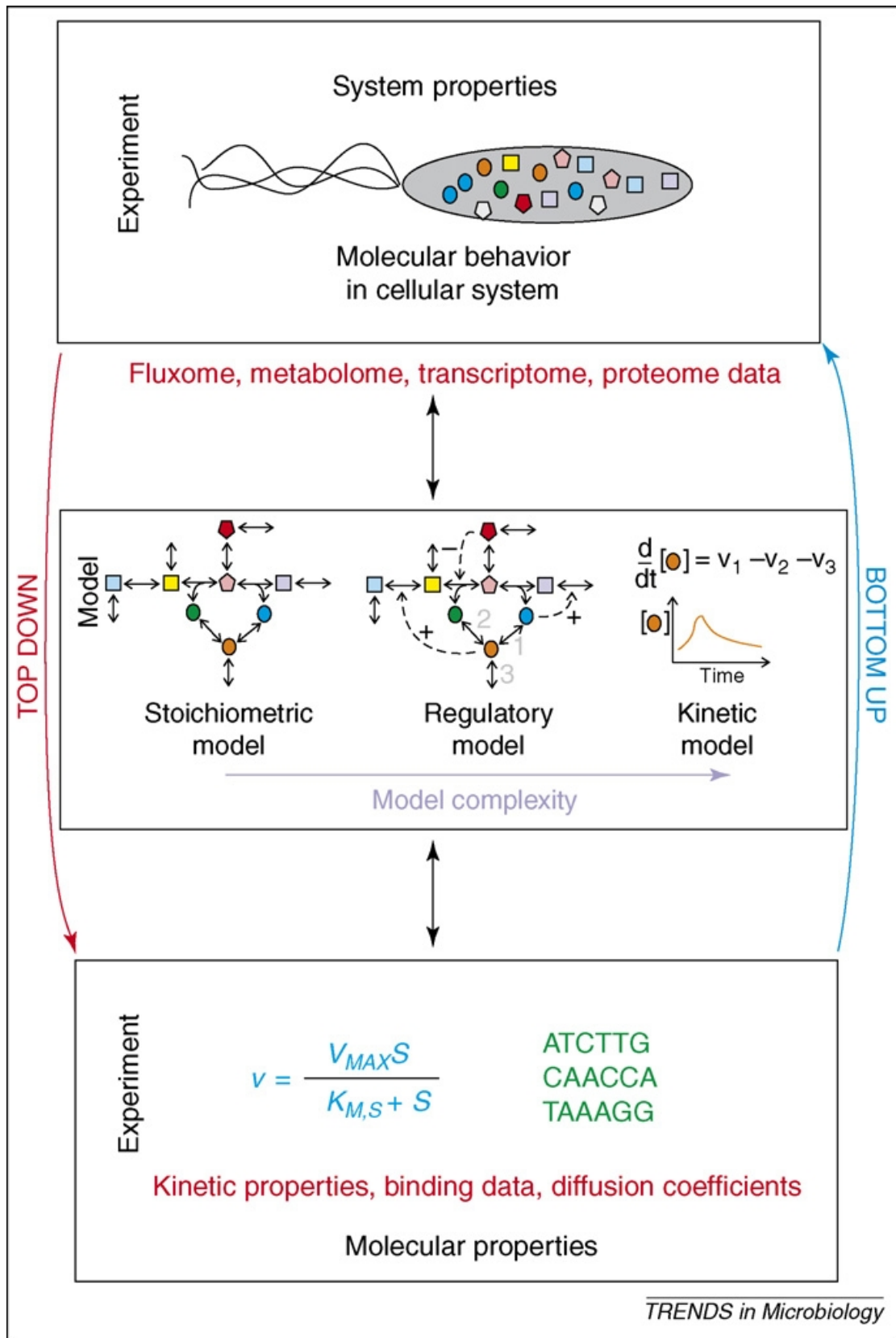


Figure 33: The schematic representation of the top-down and bottom-up approach to systems biology. (Bruggeman and Westerhoff 2007)

other supplementary methods making them indispensable. So, a qualitative analysis must be done before a quantitative (kinetic) is carried out (Runge, 2004). And it is a significant problem and major challenge for systems level studies to model the pathways usefully. For modelling it is required to identify the salient features of complex process and to construct a representation that can capture the key features while simplifying everything else which can allow focus on the essential properties without distraction. It is conceptually and computationally simple to construct the representation that captures the cause effective relationships associated with each interaction needed for pathway modelling. As suitable representations for pathways several modelling schemes have been touted such as ordinary and differential equations, stochastic schemes and Petri nets. Concentration and reaction rate data which is largely unknown for pathways and which is not essential to understand their function is required for ordinary and partial differential equation which has proven to be very successful in metabolomics that include human metabolism pathways such as cholesterol. Stochastic schemes are not efficient in describing the large number of proteins though it captures very well the individual molecular events. Even though it is conceptually complex and can approximate large flows, the formulation is very computationally simple with Petri nets (Watterson et al. 2008).

Due to lack of sufficient quantitative molecular data, systems biology could not connect to very rapidly expanding experimental molecular biosciences. For the network descriptions (**Figure 33**) i.e. stoichiometric, regulatory and kinetic models the molecular properties derived form the basis of their construction. The properties are derived from experiments that are carried out in various molecular biology laboratories and from bioinformatics. While the bottom-up systems biology starts with the molecular properties in order to construct and further to predict systemic properties that follows experimental validation and model refinement, the top-down systems biology is contrary and it is systemic data driven. Either it refines the pre-existing models that describe the measured data successfully or it starts with the experimental data to discover the pre-existing models. The kinetic models are often considered in the contemporary systems biology whereas predominant focus is on the regulatory models in top-down systems biology to analyze the data (Bruggeman and Westerhoff, 2007). It is not the first time that the concept of modeling and construction of biological pathways is discussed. Good number of works has been done and have been discussing about the construction and reconstruction of biological

pathways. Nagasaki et al. (2004) presented the Biopathway Executer that uses KEGG and BioCyc, and other databases for the reconstructs biological pathways. The tool is implemented in Java and is able to export constructed pathways into the XML format GINML of Hybrid Functional Petri net (HFPPN).

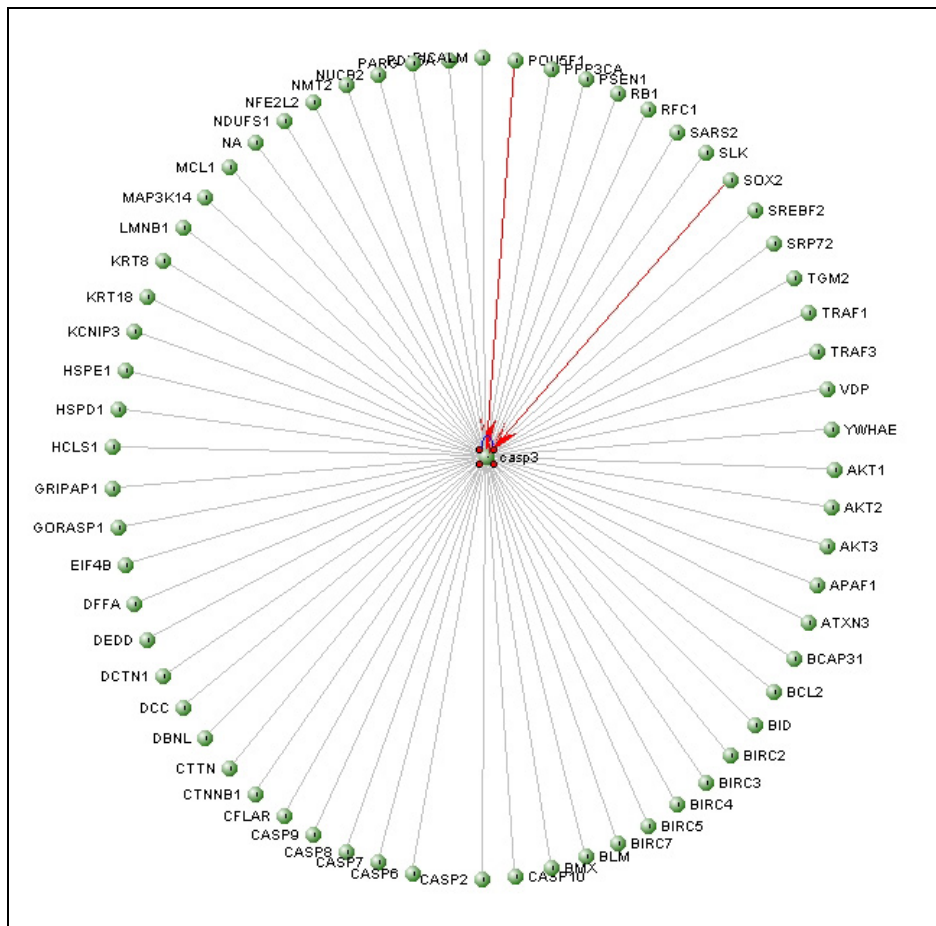


Figure 34: CASP3 gene related to cardiovascular disease and its interaction partners as shown in VisANT

VisANT 3.0 (<http://visant.bu.edu>) is developed for the visualization and modelling of biological pathways. It (**Figure 34**) is also based on Java and presented to the end users as a desktop application, web-application and also a Java applet. Along with KEGG data from different other data resources such as Predictome, SMD (Stanford Microarray Database) and GEO (Gene Expression Omnibus) was integrated to provide pathway analysis, as well as a gene and protein search. It stores the pathways in different formats i.e. VisML (VisANT Markup Language), SVG (Scalable Vector Graphic) or text format (Hu et al. 2007). Computational analysis and mathematical modelling play an essential role and they improve our capability to understand and elucidate the characteristics and functions of complex biological systems such as

metabolic, regulatory and cell signaling pathways. It is possible with the modeling and concomitant simulation to predict the cellular behavior of systems that undergo perturbations under different genetically and / or environmental conditions (Lee et al. 2009).

In signalling pathways to describe the dynamic changes in the concentration of biochemical species involved they are usually modeled by a set of ordinary differential equation. But the problem to use these quantitative data is the incomplete knowledge or non-availability of kinetic parameters. Even the existing methods for parameters estimation fail because only 30-50% of them are only known. Moreover, because of the rounding effects in the solver algorithms it is not possible to consider those numerical values of species concentration which are very small. It is not explicitly given about the possible signal flows and between species concentration e.g. feedback loops. There is more qualitative data than quantitative data became available in the last years leading to the development of qualitative methods such as discrete models and related evaluation techniques and they are used for the construction and understanding of larger reliable models as an intermediate step. Moreover in signalling pathways there are no stoichiometric equations given that govern the networks and also define the weight of the arcs. In order to model signaling pathways they have to be worked on another abstract level than modelling metabolic networks. And a generic description principle, applicable on any level of abstraction is provided by Petri nets (Grafahrend-Belau et al. 2008).

Here we propose a web-based tool MoVisPP (Chen et al. 2009) which can construct Petri net based large scale biological networks using a data warehouse approach. Furthermore, it allows the modeling and simulation of these biological networks using software suites such as Cell illustrator and allows end user interaction and network reconstruction. The tool allows the biologists and other users to construct the networks following minimal steps. It also has added features embedded in it that gives them more information about the pathway, protein, disease etc.

5.3 MoVisPP

MoVisPP (**MO**delling and **VI**sualization of **P**athways using **P**etri nets) is a web-based tool (**Figure 35**) for the construction of large scale biological networks that is supported with integrated information.

AG BI

Bioinformatics Department
Medical Informatics Department

MoVisPP

Home | Search Pathway | Search Motif | Input | Help | Contact

About MoVisPP

MoVisPP is a web-based tool to model and visualise biochemical pathways using Petri Nets.

This tool can generate Petri Nets from all pathways of the KEGG database which are available in the XML-data format. A KEGG pathway consists substantially of proteins, enzymes, compounds, drugs or glycans and there are reactions or relations between these components. While the components become places, the reactions and relations become the transitions of the generated Petri Net. Furthermore the user can choose if protein interactions from the MINT database or diseases from the OMIM database should be integrated in the Petri Net representation of the pathway. At the moment these two options are only possible if the user has chosen "Homo sapiens (human)" as organism.

The generated Petri Net is displayed as image and can also be exported in the PNML, SBML and CSML format.

The use of the tool is very easy:
At first the user has to choose a pathway that he/she wants to display as Petri Net. After the pathway selection the user can choose the organism for which the pathway should be visualised and has the opportunity to determine which reactions and relations should be displayed. In the last step before displaying the Petri Net the user can change the size of the image, the transitions and the places.

Then according to the chosen values the image of the pathway as Petri Net is generated and displayed.
If the user clicks with the mouse on the places and transitions in the image he/she gets further information of the chosen vertex.

>>> Start with the generation of a pathway [here](#) or use the link "Search" in the menu.

Figure 35: The homepage of MoVisPP along with the guidance for the user

5.3.1 Concept

For the modelling and construction of large scale biological networks it is required to know each step and each changes that happen within the network. And to construct large networks is cumbersome as it takes time and labor leading to several days for the construction of the desired networks. Also it is essential to have the kinetic parameters in place which could replicate or simulate the real world networks. And it is known it is not possible to provide the kinetic parameters for each step which is involved in the network. The key idea of this work is to construct biological networks in a qualitative manner (not dependent on kinetic data) and at the same time to construct networks using partially automated methods that avoids manual intervention which is often the case with other methods. In MoVisPP biological networks are constructed based on the concept of Petri nets using a semi-automatic approach and using integrated information from various sources that could save time and labor to a great extent. With this method it is possible to construct the desired network by extending the KEGG pathways based on Petri nets along with unique features and significant information pertaining to the protein, gene, disease etc. information which participate in the networks. Moreover the emphasis is laid on the visualization of the networks that allow differentiating the participating entities/molecules based on the nature and the associated information. Further the idea is to emphasize on the easy interaction with the system that only requires simple steps to be followed by the user to build their desired networks. Furthermore providing various export methods can allow the Petri net based large scale networks constructed by MoVisPP to be used by other applications such as Cell illustrator etc. Using software suite like Cell illustrator can allow user to import these networks and will also able to interact with the network incorporating the kinetic parameters when available and possibly able to simulate the networks.

5.3.2 Architecture

MoVisPP has multi- tier architecture (**Figure 36**). It has a graphical user interface on the top that allows client interaction using a common web browser. Lying below is the application and logic layer connected to the data warehouse that runs on the server side. Having implemented in Java the application logic is platform independent. For

facilitating the web application MySQL is used as a database server. A Java Database Connectivity (JDBC) driver for MySQL facilitates access to the database from the

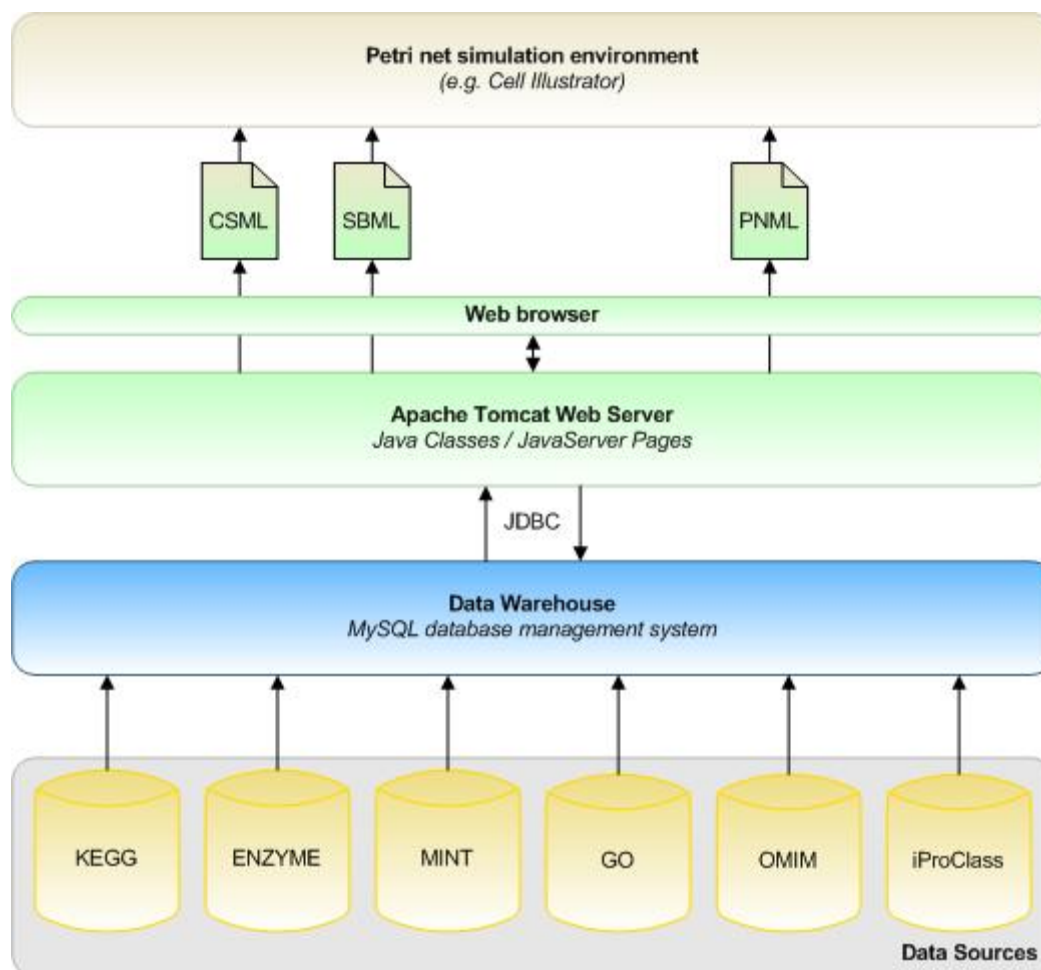


Figure 36: A detailed system architecture of MoVisPP describing the various layers and their action.

Java classes. JavaServer Pages (JSP) contains Java code that can be included via JSP elements in static HTML code. Based on the end user activity, the static HTML parts can be dynamically complemented. Whenever a JavaServer page is opened in a web browser, a request is sent from the client to the server. The web container (also called servlet container) from the server creates, compiles and /or executes a Java servlet. A response in form of a generated static HTML website is sent to the client from the server. MoVisPP uses an Apache Tomcat 6.0.2 servlet container to generate HTML code.

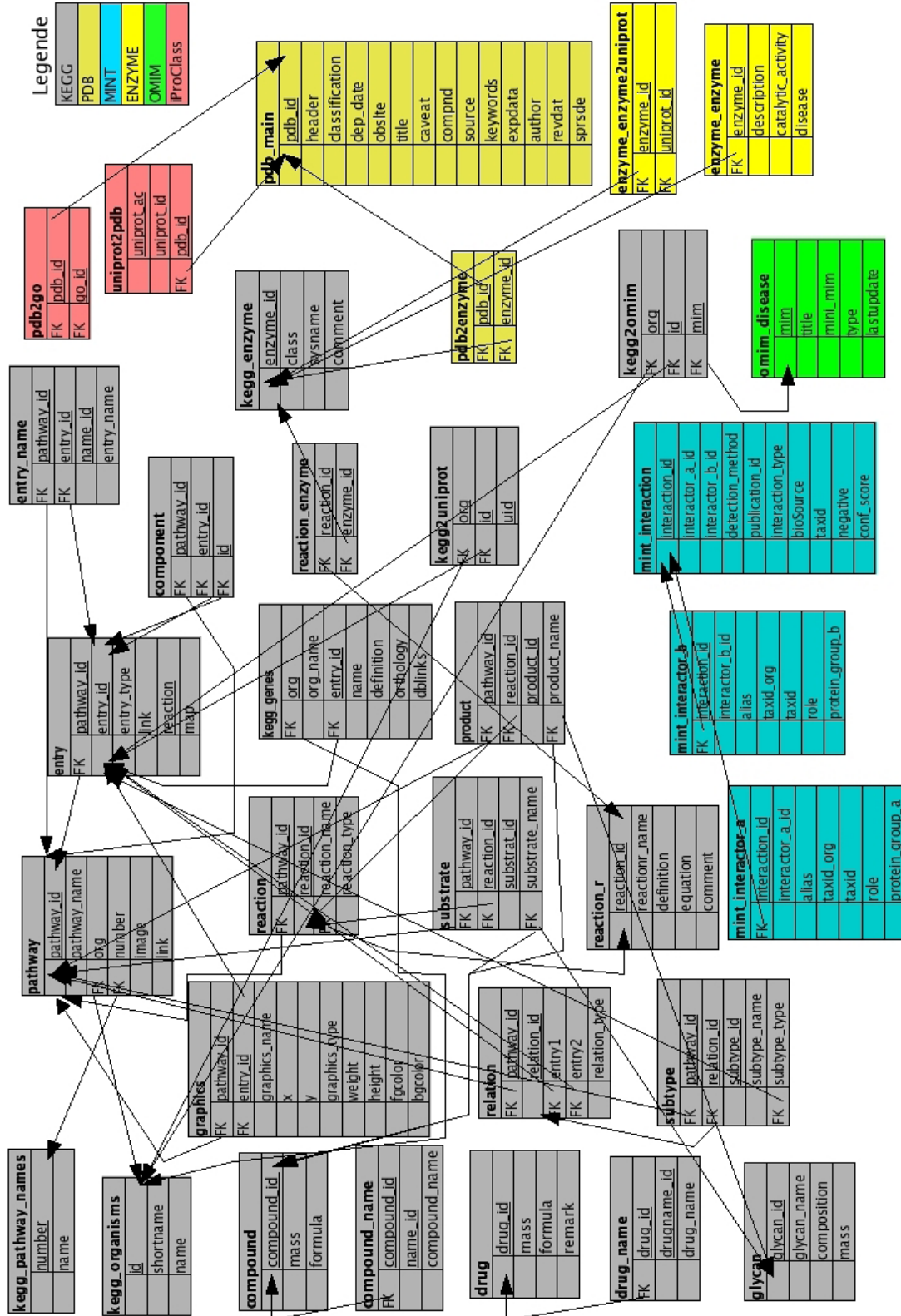


Figure 37: The graphical representation of the data warehouse structure (Spangardt, 2007)

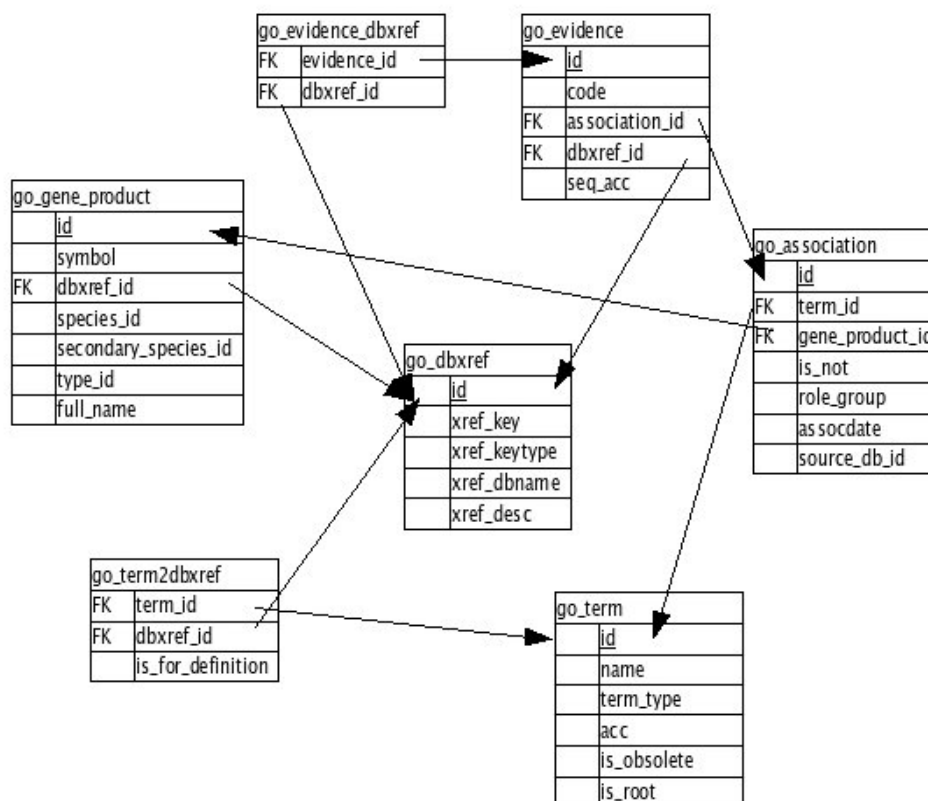


Figure 38: The figure shows the relationship between the integrated data from GO (Spangardt, 2007)

The web application being user friendly allows them to choose from the listed options to proceed further with the modelling and visualization of biological pathways. Following a few steps it will be able to generate the pathways as Petri nets. First, the end user must select one from the listed KEGG pathways followed by a specific organism. Based on the users selection the server will decide about the reactions and relations from KEGG and also the interactions from MINT along with the OMIM data allowing modelling and visualization. The size of the transitions, place and the entire image can be set by the user. Following these steps and based on the user selection the generated pathways as Petri net is displayed as an image in a common web browser. With the image map functionality the end user will be able to obtain the detailed information about the various parts of the Petri net.

The detailed data warehouse scheme shown (**Figure 37**) depicts the relationship between each entity and the integrated data. The data stored in the data warehouse is cross referenced to each other via unique id. The scheme shows the different tables along with their unique id (primary key) and how they are connected to each other

using different data entity relationships using the key or other information. Each table shows the domain information that is integrated from different sources. The interconnection shows how the information present in each data source is unique and how much they share with other sources. Further it is shown in (**Figure 38**) how the GO terms are referred in the underlying database.

5.3.3 Implementation

MoVisPP is supported with a data warehouse at the backend containing the information from different sources such as Enzyme (Bairoch 2000), MINT (Chattraryamontri et al. 2007), OMIM (Hamosh et al. 2005), PDB (Berman et al. 2000), iProClass (Huang et al. 2003), KEGG (Kanehisa, 2000), GO (Gene Ontology Consortium 2006). A relational database management system at the backend contains all necessary information that is required for the visualization of the pathways. Java-based parsers are written for data integration that carries out the job semi-automatically and the merging and transformation of the data is carried out manually. For the visualization of pathways as Petri nets JUNG (Java Universal Network/Graph Framework) is used. For allowing end user interaction using common web browser, a web-based graphical interface has been implemented using JavaServer Pages (JSP). With this approach it is possible to the networks in MoVisPP with two steps approach and it is platform independent as it is implemented in Java. In MoVisPP the networks are created on the fly along with the integrated related information.

Furthermore the tool can export the networks to different formats such as Systems Biology Markup Language (SBML), Petri Net Markup Language (PNML) or Cell System Markup Language (CSML). Moreover the exported network can be modelled and simulated using some of the Petri based software suites e.g. Cell illustrator that allow further interaction. For this work Cell illustrator is used since it is commercially available and we have the license to access. Moreover the tool is highly standardized and reliable and has been used by many researchers and has added features compared to other Petri net tools. Cell illustrator handles the different file formats that are mentioned above with ease and also allows the user to store in the above formats which allow the exchange of files across different applications.

Like other tools MoVisPP follows certain basic steps for the modelling and visualization of the integrated data. It uses the data from KEGG to model and visualize the pathways as Petri nets. The state is represented by the “Place” and

activity of the system by “transition”. The generated graphs follow the Petri net model where the places are represented by circles and the transition by a rectangle. The arcs represent the dependencies. It follows a qualitative modelling approach i.e. non-availability of the biochemical parameters. The genes, proteins, enzymes, chemical compounds, glycans, pathways and the active components are shown as circles. It is possible for the user to provide the kinetic and other biochemical parameters to the participating places and transitions either by referring to the literature and other related works further by exporting the generated pathway into other simulation environment such as Cell Illustrator.

Each entity is given a unique color Proteins – blue colored places and the presence of gene information by providing a thick red border. Also, the information about the various enzymes, proteins and genes is also provided. By this approach it is useful to understand and give an overview of those details that are missing e.g. gene expression data. It is essential to know about the information of the participating enzymes, substrates and products to model the pathway. For this information from the KEGG database is directly incorporated. But, to give nicer view to the user some of the compounds are not visualized assuming these components are ubiquitous and they participate in every reaction. Also, the co-enzymes are not visualized. Enzymes are catalysts and they are not consumed in a reaction. So, a place which represents the enzyme and a transition representing the reaction is connected by an arc that is drawn in both the directions. In case, if the reaction is reversible in the model the arc is drawn from the “substrate” that is depicted as a place to the “reaction” depicted by a transition and also the “product” again represented by a place and vice-versa.

Just moving away from the conventional Petri net model and representation in MoVisPP, a reaction i.e. given by R-number from KEGG is used as a label on the arc which is between the product or enzyme and the transition and it is shown in the **Figure 39**. And this does not relate anything with the weighted arc or any other mathematical function but it is only to present a better understanding. According to KEGG there are several kinds of relations (<http://www.genome.jp/kegg/docs/xml/>) such as GEl- interaction within gene expression, PPre- Protein – Protein interaction, PCrel – protein and chemical compound interaction, ECre- enzyme – enzyme relation, Maplink – link to the pathway map.

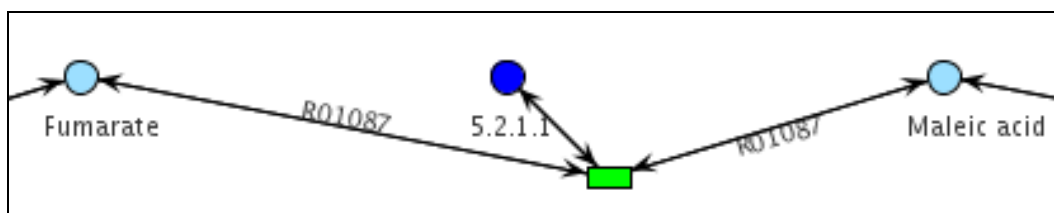


Figure 39: An enzymatic reaction step generated by MoVisPP

In MoVisPP, the PCrel shows arc in one direction from a pre-condition through a transition which is colored red. In order to show which products of the current pathway are processed or also present in another pathway and these relations are a sub-class of a compound, shown by the connection of Maplinks. Ecrels – the enzyme-enzyme relations are always connected via chemical compounds and the enzymes are not consumed always and each place has arc in both the directions as shown in **Figure 40**. Moreover there is another arc that traverses from the place of the chemical

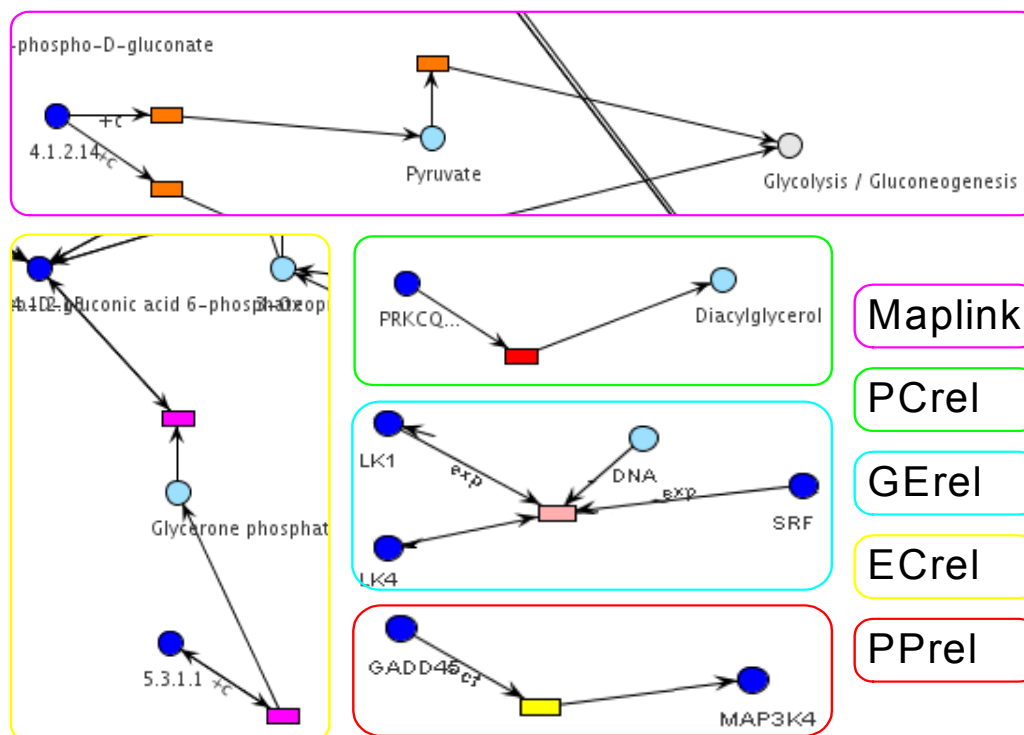


Figure 40: Graphical representation of different KEGG relations as a Petri net model generated by MoVisPP

compound. Protein-protein interactions are modeled from the pre-condition to the post-condition via a transition. At the same time the presence of an intermediate compound will be shown using additional pre-condition and post-condition into the model. An inhibitory arc which is marked with a circle is a salient feature and is at the end of an arrow of the transition side. In the Petri net model PPrels are colored

yellow. GErels i.e. gene expression relations which is colored red, has three sub-domains the expression, the repression and indirect effects. They are connected by a transition in principle and the presence of genetic information such as DNA is shown as a connection to the transition and indicated by a circle by the end of the circle are repressions.

The integrated information from OMIM is represented as green places and in the Petri net models they are connected as light-yellow transitions. The information that is correlated between the diseases and the genes / enzymes is available in OMIM. Due to the lack of information it is not possible to illustrate in a detailed manner of the dependencies between this data. This persuaded to the projection of the essential information in a simplified manner in the Petri net models. In MoVisPP, it is also possible to visualize the protein-protein information from MINT and only mammalian interaction data is only integrated. Here also the simplification of the model is followed during visualization of the interaction where “A” protein is connected to the “B” protein via a transition.

5.3.3 Application

The conventional way of constructing pathways using other Petri net tools are more time-consuming and manually intensive, whereas the semi-automatic method can build a Petri net based network along with more integrated information comparatively at a faster rate and less time. Furthermore, this method also provides the opportunity to simulate the pathways using commercial and non-commercial Petri net-based tools. Thus, this approach has an edge over the currently available methods of constructing pathways.

The KEGG and other databases contain lots of information about different pathways that are associated with several diseases. In KEGG these pathways are projected as maps that show the connection between the protein, gene, enzyme etc. along with the reactions involved. If a KEGG pathway (**Figure 41**) showing number of cellular signals initiate activation of apoptosis in different ways that depend on the cell type and their biological states is supported with the extra features mentioned above, it will always have an advantage. By providing the user with a Petri net version of the KEGG based network along with other information can always be useful for future study and large scale modelling of pathways and atlas.

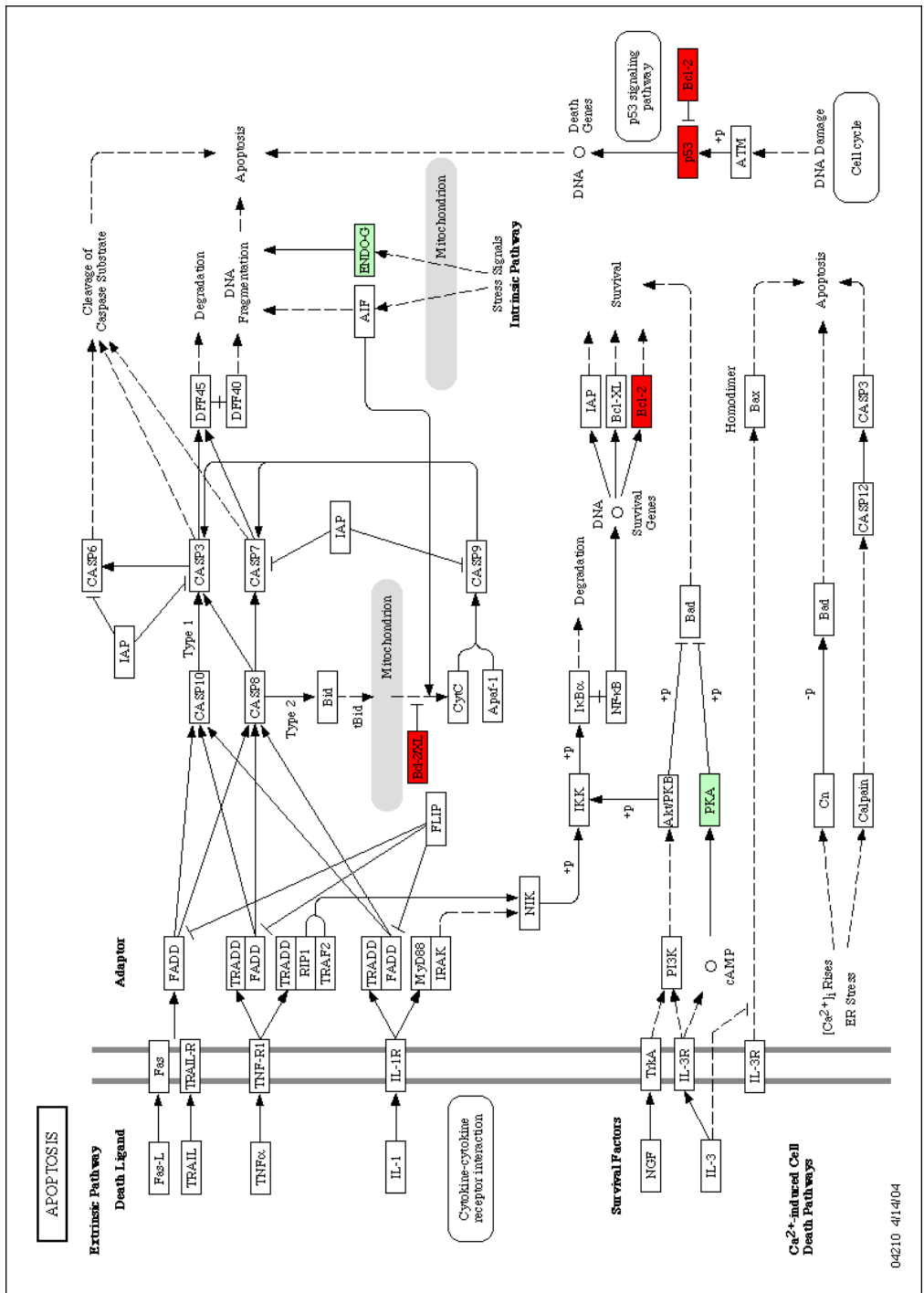


Figure 41: The KEGG (<http://www.genome.jp/kegg/>) representation of the apoptosis signal pathway explains the role of Bcl2 and p53 showing the different ways of activation of apoptosis. While the crossbar arrowheads indicate inhibition the branching arcs go to alternative as well as to concurrent successors.

The screenshot displays the MoVisPP web interface. At the top, the MoVisPP logo is visible. The main content area is divided into two sections: 'Step 1: Selection of the pathway' and 'Step 2: Selection of the details'. Step 1 includes instructions on how to use the KEGG identifier and a search bar. Step 2 includes a dropdown menu for selecting the organism (Homo sapiens) and various checkboxes for filtering the map. A large network diagram is shown in the background, and a legend is overlaid on the right side of the diagram. The legend lists various biological entities and their status (shown or not shown).

Step 1: Selection of the pathway

If you don't know the KEGG identifier of the pathway you would like to visualize as Petri Net, use the first form to choose the pathway and in the next step you can choose the organism and other details. Otherwise use the second form and enter the KEGG identifier in the text fields. In the next step you can choose the other details.

Please select the pathway that should be visualised.

Apoptosis (04210)

Step 2: Selection of the details

In the last step you have chosen the following details:

Please select the organism for which to visualize the pathway:

Please select if and which kind of reaction to visualize:

Reactions

Reactions

Organism specific reactions

All reactions available for the Reference

Relations

Organism specific relations

All relations available for the Reference

Protein-protein interactions

Protein-compound relations

Enzyme-enzyme relations

Gene-expression interactions

Links to other maps

Please choose if you want to visualize the pathway:

Diseases from OMDM database

Please choose if you want to visualize the pathway:

Protein interactions from MINT interaction database

Legend

Pathway:	Apoptosis (04210)
Organism:	Homo sapiens (human) (hsa)
Proteins/Enzymes:	shown, not shown , gene information available
Compounds:	shown
Glycans:	shown
Drugs:	shown
Diseases:	shown
Reactions:	Organism specific
Relations:	Organism specific
	ECrels: shown
	PCrels: shown
	PPrels: shown
	GERels: shown
	Maplinks: shown
	DisRel: shown
	MINTint: not shown

Figure 42: A screenshot of MoVisPP showing its user friendly browser in the background. The user can interact with the system for generating their desired map by following few steps involved. Also, in the forefront is a qualitative model of

apoptosis pathway generated by MoVisPP showing the network along with its supported legend.

For example, gene and protein expression pattern studies of Bcl2 and Tp53 show their role in cardio-vascular diseases. Also, these genes play a major role in other neurodegenerative diseases and in apoptosis. While a search for the gene Bcl2 in VINEdb described in the chapter 2 presented an interactive composite network (**Figure 23**) along with other information. A search in the KEGG resource yielded a list of the pathways in which they are involved. An apoptosis map generated from KEGG (**Figure 41**) indicates the participation of these genes Bcl2 and Tp53. But this map is static and has links to the information available within the KEGG database.

As mentioned earlier MoVisPP has a user friendly approach (**Figure 42**) for extending the biological networks. It guides the user and allows them to choose from the options listed leading to the construction of their desired network. **Figure 42** shows the apoptosis network generated by the MoVisPP system choosing from the listed options involving pathway selection, details of the organism, reactions, relations, disease and about protein interaction along with the stretch factor and size. Unlike the conventional modelling methods which require quantitative data (i.e. kinetic parameters etc.) the networks generated by this method are not dependent on the quantitative data. In a way it allows the qualitative modelling using a Petri net based approach. Moreover it is not possible to provide all the biochemical parameters for all the reactions and other biochemical data dependent units that participate in the pathways Furthermore, the networks created from MoVisPP can be exported and used by several other applications that support Petri nets e.g. Cell Illustrator (<http://www.genomicobject.net/member3/index.html>).

Figure 43 and **Figure 44** are snapshots of a portion of the KEGG pathway-based apoptosis network generated by MoVisPP and are created at run time. These interactive networks are supported with some more information available from the databases integrated such as Enzyme, MINT, OMIM, PDB, iProClass, GO and KEGG. As described in the earlier sections of this chapter the network is supported with a legend that has color indicators and explanation for each entity involved along with the detail of the organism.

The generated Petri net based networks along with the added information of enzyme, gene, protein etc. can be exported by following simple steps into different formats

using the built in export function that provides different possibilities to generate, i.e. CSML, PNML and SBML based outputs in order to be used in other software suites.

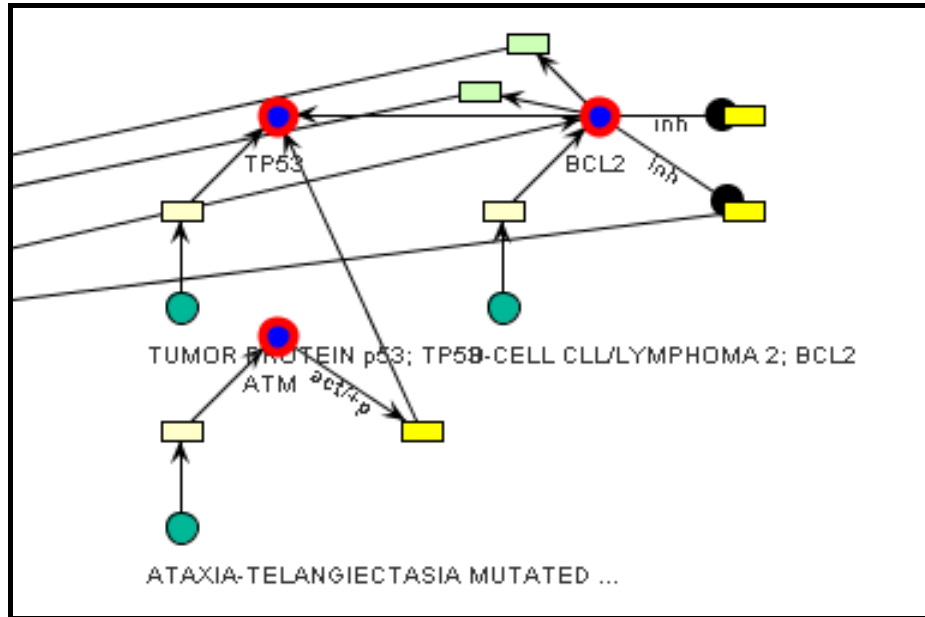


Figure 43: Enlarged screenshot of the MoVisPP generated qualitative model of apoptosis pathway that is shown in the inset of Figure 40. This image is supported with integrated information from the data sources and shows the presence of the genes Bcl2 and Tp53. Every participating entity is given a unique color by the system.

The exported network can be further interacted by allowing them to be imported into Petri net based software suites such as Cell Illustrator which allow their simulation making them more dynamic and user friendly. This will allow the end user to further expand and incorporate the quantitative data such as the biochemical parameters for the participated reaction, enzymes etc. Thus the Petri net based apoptosis network (**Figure 44**) can be modeled or upgraded into a quantitative network further allowing its simulation.

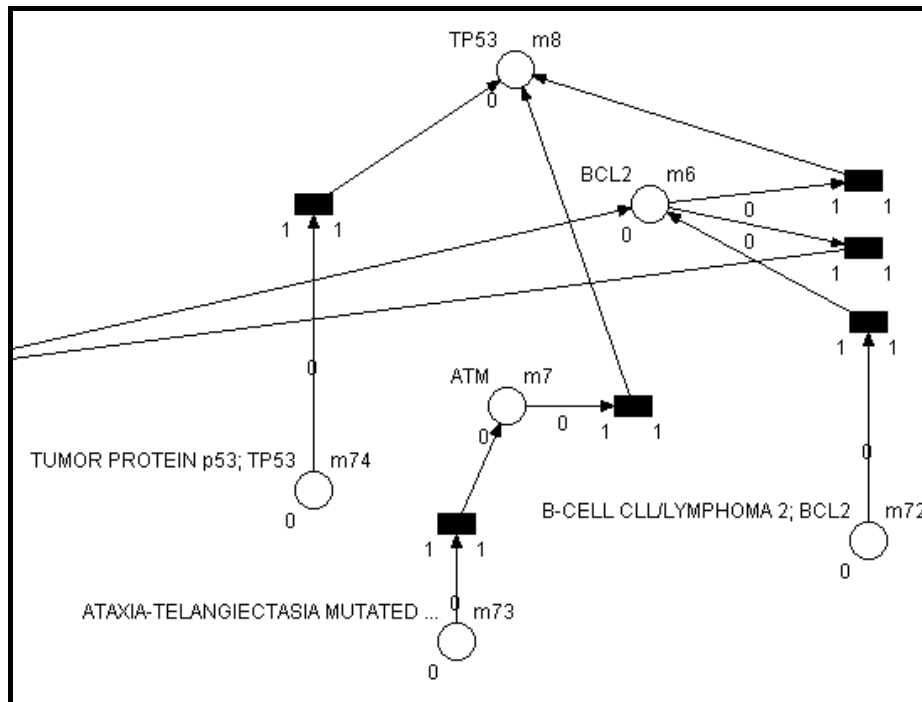


Figure 44: A part of the Cell Illustrator image that is exported from MoVisPP and it is based on the qualitative model of apoptosis pathway.

5.4 Summary

It is not possible to provide the quantitative data for all the participating reactions in the large scale systems or biological processes. Moreover these quantitative data also has restrictions for their use in the algorithms. In the recent past the generation of qualitative data is more compared to the quantitative data. This led to the explosion of several qualitative methods and also modelling of biological networks. This work is an approach in the direction of large scale qualitative network construction. The large scale construction of biological networks proposed throws yet another dimension into the modelling and simulation of pathways / networks at the same time allowing a higher level approach. By providing the backend data warehouse with different sources it has become possible to bring in diverse and related information of the molecule or the entity involved in these pathways. Moreover by giving the end users an option to construct pathways using different Petri net based suites it allows further interaction of the pathways of their interest by introducing the required parameters during the modelling process using their friendly software suites. The network can be expanded and applied to CardioWorkBech and similar projects by integrating protein-protein interaction results which will be discussed in the next chapter.

MoVisPP is available at <http://agbi.techfak.uni-bielefeld.de/movispp/>.

5.5 Outlook

The large scale construction of biological networks by MoVisPP opens new vistas in the area of modelling and simulation and in systems biology. The present underlying database which supports MoVisPP can be extended in order to accommodate biochemical / kinetic and other parameters also if they are available. The visualization method can be thought in different aspects if possible and also the export functions and formats. Even protein-protein interaction (PPI) results can be incorporated that can bind the multifarious biological data and could still present a better overview of the networks. Also, by combining the PPI networks with the related information can give more information about the relation and reaction in concert with the integrated molecular information pertaining to the participating proteins, pathways, enzymes etc.

The following chapter deals with the significant information that deals with the method to construct networks in a unified manner. This method will bind diverse information along with PPI information so that it will be able to learn the behavior of the proteins, pathways etc. in unison.

Chapter 6

Application

6.1 Construction and reconstruction of biological networks – an integrative approach

“If we hope to understand biology, instead of looking at one little protein at a time, which is not how biology works, we will need to understand the integration of thousands of proteins in a dynamically changing environment. A computer will be the biologist's number one tool” said Craig Venter. When viewed through the lens of network biology valuable insights into the action of drugs and for knowing the ways to improve the drug discovery for complex diseases are provided. Network biology can also play a significant role in identifying drug targets. “Polypharmacology” appreciated in the recent years is a phenomenon that involve many effective drugs in different therapeutic areas such as oncology, psychiatry and anti-infectives acting on multiple targets rather than single targets. Not only the drug targets that are commonly involved in multiple diseases but also drugs that commonly act on multiple targets are revealed when the polypharmacology network is mapped onto to the human disease-gene network. More than 40% of drug targets which are mapped with disease genes are mapped to several diseases. The studies and analysis of OMIM database that consists of genetic association reveals, *most genetic diseases share their genetic origin with several other diseases* (Hopkins AL, 2007).

“Diseases such as atherosclerosis and hypertension comprise a diversity of different disease subtypes involving multiple organs and tissue types. Operating within each tissue (and each cell within a given tissue) are a number of molecular networks that ultimately drive the onset of disease. These networks are context specific and sensitive to internal and external environmental conditions as well as genetic background. Variations in the connectivity structure of these networks are induced by variations in the genetic background and environmental conditions, where these variations in turn lead to phenotypic variations, including disease. Studying the molecular networks in all relevant tissues and associating them with clinically relevant phenotype data to identify the networks driving disease are among the goals

of systems biology applied to disease research. By taking a more holistic approach, it may be possible to better understand the complex interplay among tissues, molecular networks, and environment that leads to disease” (Schadt and Lum, 2006).

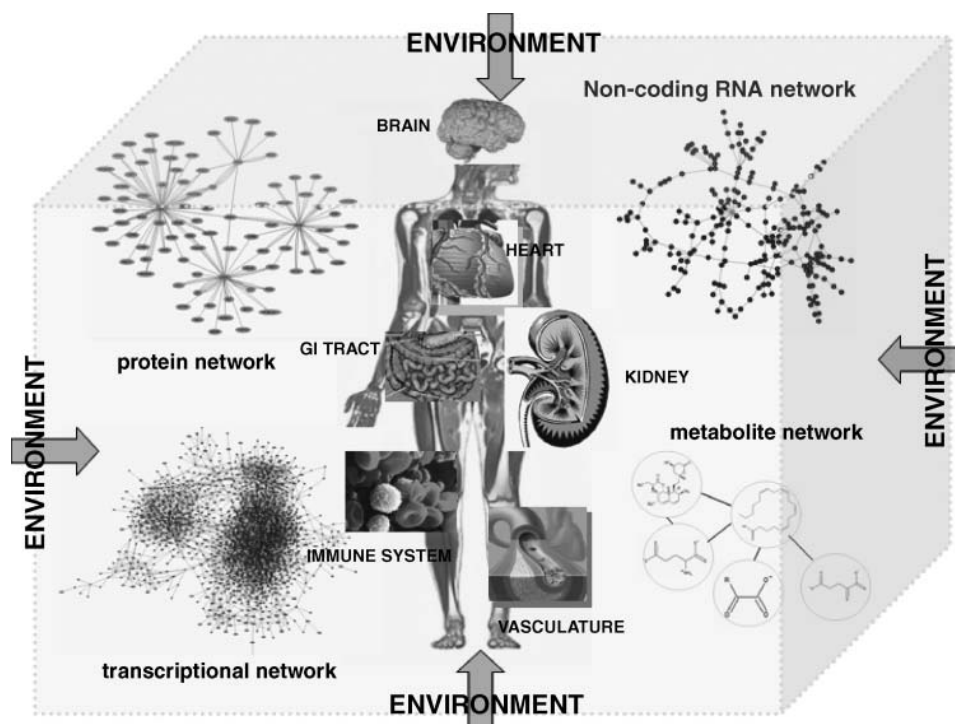


Figure 45: Schematic representation of diversity of networks in tissues. (Schadt and Lum, 2006)

In any given tissue the functioning of the molecular networks is diverse (Figure 45) that include genomic networks, protein interaction networks, protein state networks, networks of coding and non-coding RNA, signalling networks and networks of metabolites. Acting in concert not in isolation within each cell, these networks form complex giant molecular networks interacting with each other within and between the cells driving all the activity among the tissues along with signalling between tissues. Complicated physiological processes are induced that can manifest as disease due to the variations in DNA and environment that further leads to the changes in these molecular networks (Schadt and Lum, 2006).

Signalling pathways are complex (Figure 46) and can be compared to an integrated circuit. The signalling pathways like the integrated circuit perform complex calculations or vice versa that processes and transfers the information across the cells and systems in order to achieve a certain function. Starting with a small stimulus the

signal transduction becomes “signal cascade” due to the increase and participating of several proteins and other molecules in the process.

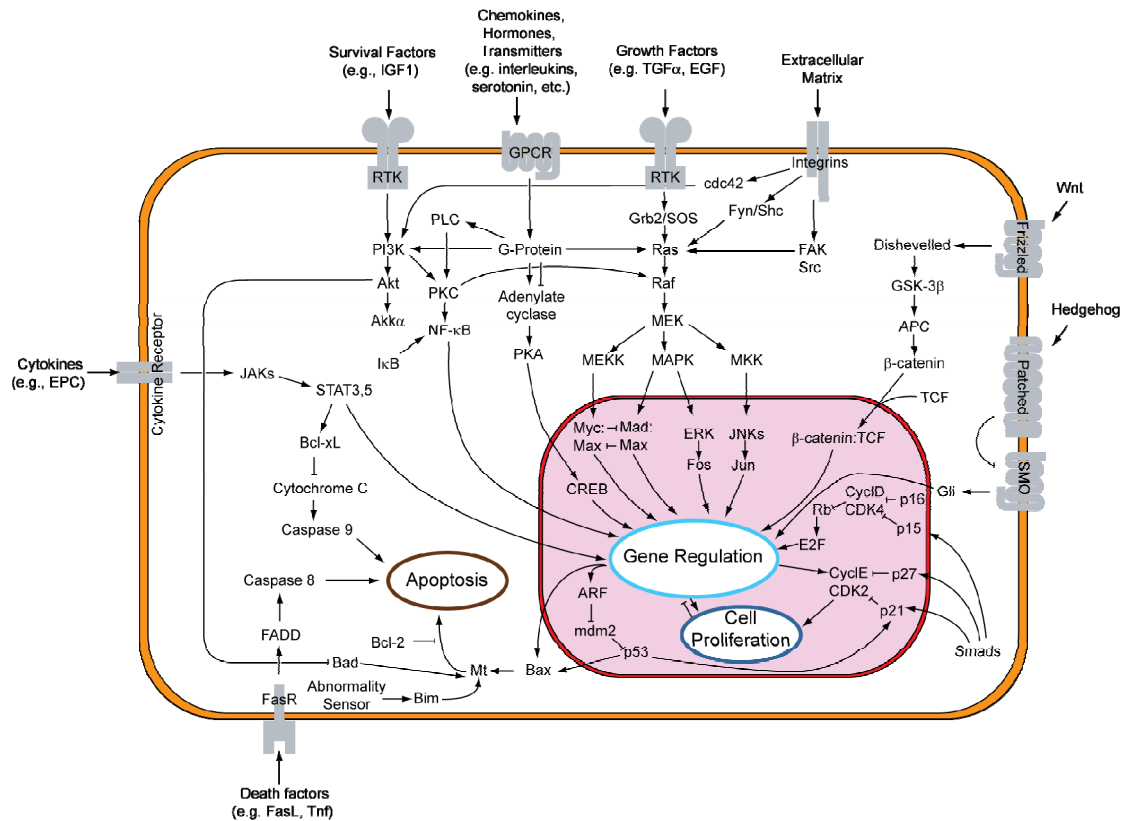


Figure 46: An overview of signalling pathways (Source: Wikipedia- Lodish H, 2003).

Many signalling pathways respond to stress and among them Mitogen Activated Protein Kinase (MAPK) family members are crucial. It has been deemed that MAPKs are stress responsive and they are involved in apoptosis. Then it was initially thought the regulation of apoptosis by MAPK is more complex and also quite controversial (Wada et al. 2004). MAPK family members are involved and regulate large variety of cellular processes such as cell growth, development, differentiation, cell cycle, death and survival. It has been identified in the mammalian myocardium there are several MAPK subfamilies involved with unique signalling pathway apparently for each of them. Moreover, the cascades do not have similar processes they differ in their sequence of upstream activation and also in their substrate specificity downstream. It is found in the heart, the activation of MAPKs family has a significant role in the pathogenesis of various processes e.g. in ischemic and reperfusion injury, in the cardioprotection conferred by ischemia-or pharmacologically-induced preconditioning and in myocardial hypertrophy and its later transition to failure of the heart (Ravingerova et al. 2003).

Investigators have been researching for many years different methods of preconditioning the myocardium to prevent from damage induced by ischemia. Human myocardium loses its capacity to tolerate and also fail to respond to various forms of stress. These changes are noticed in normal ageing. With rise in the individual's age, complications also increase and it is likely they may experience an ischaemic stress and other cardiovascular complications. With cardioprotective treatments aged population will be benefitted. Several methods including the exercises, heat stress, oxidative stress etc. provide cardioprotection or preconditioning. Though it is not clear whether the aged myocardium will be able to adapt to these preconditioning stimulus, it has been observed there are several alterations that take place in the activation and expression of the key proteins. These proteins include the sodium-hydrogen exchanger (NHE), heat shock protein 70 (HSP70), nitric oxide synthase (NOS) and the mitogen-activated protein (MAP) kinases i.e. extracellular signal-regulated kinase (ERK), c-Jun N-terminal Kinase (JNK) and p38 (Taylor and Starnes, 2003).

Mitogen activated protein kinases (MAPK) are serine-threonine protein kinases. They are involved in numerous processes (**Figure 47**) that are important to cardiac surgery such as vascular permeability, vasomotor function, cytokine production and reperfusion injury. MAPK are expressed in multiple cell types which include vascular endothelial cells, cardiomyocytes and vascular smooth muscle cells. They also function in cellular signal transduction cascades and are being activated by a diverse range of stimuli including ischemia, vasoactive agents and shear stress. C-Jun NH₂-terminal protein kinases and p38 kinases, extracellular signal-regulated kinases are the three major identified MAPK families. And extensive investigation has established the roles for these three families in cardiovascular signal transduction pathways. Activity of these signal cascades may contribute to the myocardial reperfusion injury and increased pulmonary vascular permeability which is observed after cardiac surgery with cardioplegia and cardiopulmonary bypass (Khan et al. 2004). Also MAPK plays a key role in intracellular signal transduction and regulation among the 518 recognized protein kinases which are recognized in the human kinome. Classic MAP kinases that are implicated in a wide range of cellular processes from cell growth and proliferation to apoptosis are extracellular signal-regulated kinases

(ERK1/2), p38 kinase ($\alpha, \beta, \lambda, \delta$), c-Jun N-terminal kinases (JNK 1, 2 and 3) and big MAP kinase (BMK or ERK5). Specific functions of MAPK in heart have been a focus

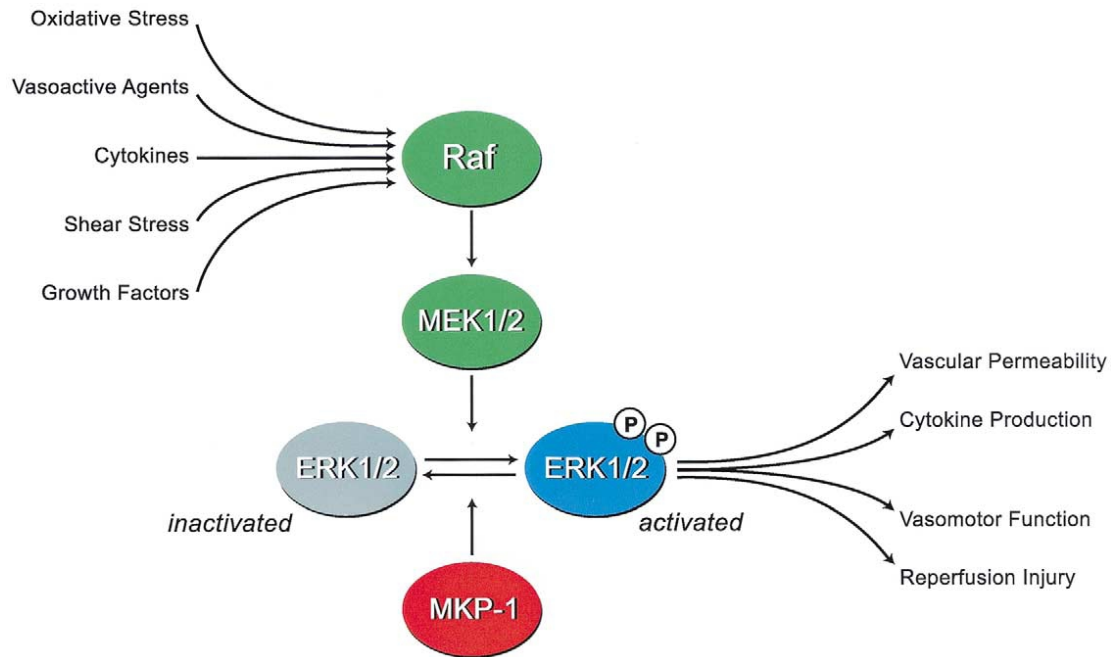


Figure 47: Activation of MAPK and its downstream effects (Khan et al. 2004).

of intensive study as they are ubiquitously expressed. MAPK activation is observed at different stages of heart disease progression that include hypertrophic cardiomyopathy, dilated cardiomyopathy and ischemia / reperfusion injury other than recognized role in cardiac development in the recent past both in human and animal models. Significant insights into the regulatory mechanisms and potential downstream targets of MAP kinases in the heart have been revealed from the recent molecular studies. The animal models have begun to yield evidence for the *in vivo* function in cardiac development, physiology and pathology when their MAP kinase pathways are genetically manipulated. For targeting specific MAP kinase pathways on cardiac function in cellular and animal models with high impact, improved pharmacological agents with high potency and specificity will be a major help. From the above discussed advances it is evident that MAP kinases are significant players in cardiac physiology and pathology (Wang, 2007).

Apoptosis is a very active process that leads to cell death. With the activation of variety of extracellular and intracellular stimuli the signalling pathways are initiated which in turn mediates the process, apoptosis. Apoptosis has a very important role in normal physiological processes comprising of functional self-organization processes

in the immune system and central nervous system, the changes that occur morphogenetically during the embryonic development, tissue homeostasis in adult animals and in damaged cells removal. Also, it is highly involved in many human disorders pathogenesis such as cancer, AIDS and also other immune disorders, many neurodegenerative diseases including Parkinson's disease, Alzheimer's disease, ischemia, stroke and also in cardiovascular diseases. In the last decade a large number of cellular factors that are associated with apoptotic signalling pathways have been identified. These include cell cycle regulators (pRb and Cdk inhibitory proteins), Bcl2 family members, cell surface receptors (death receptors of the tumor necrosis factor receptor family), the inhibitors of apoptosis protein (IAPs) and many cell adhesion proteins and stress-response proteins such as heat shock proteins (Cho and Choi, 2002).

The following **Figure 48** shows the MoVisPP generated human apoptosis pathway along with the integrated information. As discussed in the previous chapters the network generated is based on the concept of Petri net based approach generated on the fly. Moreover the network is supported with more information. The information is from different sources that give the detail of the protein, gene, disease etc. involved in the pathway. It is well known the apoptosis pathway is again linked to different pathways and variety of other information and it is a part of a greater biological network or a global biological map. This network when combined with other approaches along with more information can be more useful and can give newer insights into the understanding of network behaviour and protein / gene relationships.

“Systems pharmacology is an emerging field that uses both experiments and computation to develop an understanding of drug action across multiple scales of complexity ranging from molecular and cellular levels to tissue and organism levels. By integrating multi-faceted approaches, systems pharmacology can provide mechanistic understanding of both the therapeutic and adverse effects of drugs. This includes understanding of how drugs act in different tissues and cell types, as well as the issues of multiple actions within a single cell type due to the presence of several interacting pathways. Such studies are important from a translational perspective because they help identify new drug targets. A general understanding of drug action requires a systems level view rooted in the human genome. Implicit in such understanding of drug action is also the knowledge of how complex diseases originate

in the context of the whole genome of an individual. This type of understanding will come from various sources of data such as physiological, biochemical and genomic parameters. Integrating these datasets requires an array of computational approaches. One particularly valuable approach is the use of network analysis of cellular systems.” (Berger and Iyengar, 2009)

With the availability and ever increasing volume of molecular data enhances our ability to understand the capacity of study cell behaviour. Further to explore and exploit molecular data, one must also investigate the various links between genes and proteins; the relationship between protein structure and function and the united effects of many proteins acting on and their interactions with the mixture of small and large molecules within a cell. This help in the study of gene regulation and metabolic pathways. Each of these molecular data must be stored in different databases and analyzed. Many database systems offer access via the internet which stores information about genes and proteins (EMBL, UniProt, PIR etc.). This technique allows the analysis of metabolic processes in order to understand the molecular logic of cells. Hence there is a need for the important methods of modelling and simulation which influences the domains of medicine and genetics (human) at a microscopic level (Collado-Vides et al. 1999).

These studies put forth various hypotheses and the research results enhance the knowledge of drug design, disease and networks at a larger scale. In this study the rudimentary biochemical data as mentioned in the previous chapters is generated by the experimentalists /scientists who are participating in the EU Project CardioWorkBench - *Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analyses*. The biochemical data are generated with the goal to design drugs for the cardiovascular diseases. Using the raw data or the experimental data the idea is to use them in the tools that are discussed in the previous chapters and also trying to extend the results in order to answer some of the questions mentioned before. Here genes such as Bcl2, Rac1, Birc6, Cdc42 which are found to be the key players are considered for the analysis. By this method diverse information regarding the genes are generated by the in-house developed tools (VINEdb, MoVisPP) and also by using some other tools. Interesting results generated are further analyzed and then the idea to present the multifarious data as a integrated or unified network is proposed.

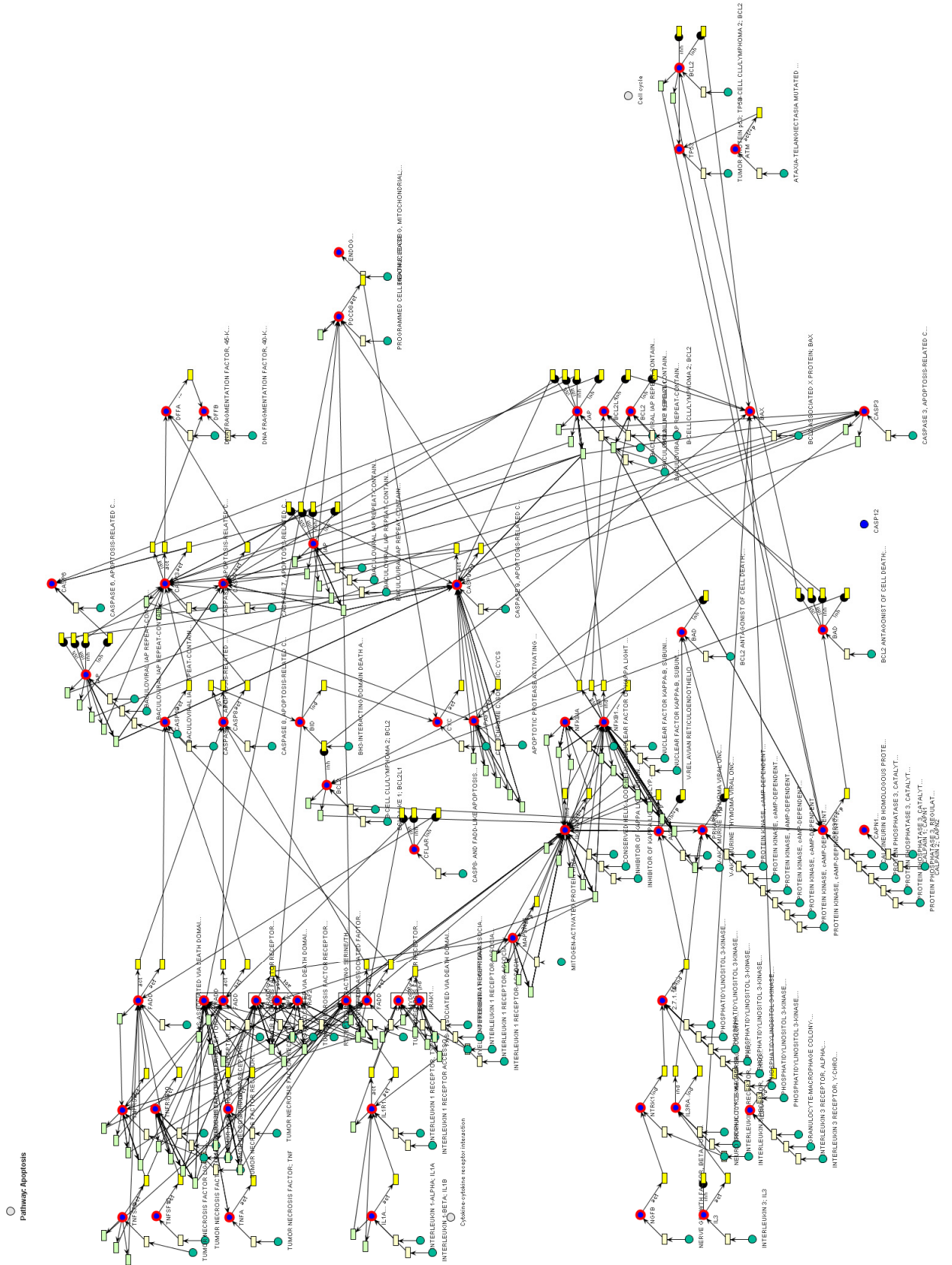


Figure 48: Human apoptosis pathway generated by MoVisPP along with the integrated information.

The output for Bcl2 in human yielded a network (**Figure 23**) that show the various domain information such as GO, disease, interaction associated with Bcl2 from the data warehouse. The results from VINEdb are further expanded for network construction / reconstruction using an integrative approach that allows network growth. Using a combination of manual and automated methods it is possible to re / construct a biological network integrated with more molecular data information that can be in parallel with other results. Furthermore with this network it is possible to decipher and relate more information in context with the behaviour of the protein.

At the same time MoVisPP yielded apoptotic pathway (**Figure 48**) along with other information from different sources. As mentioned MoVisPP constructs Petri net based network along with the integrated information from different sources. The different networks produced by the tools are considered and it paved a new direction for the analysis and reconstruction of the biological network. Furthermore the search result for the Bcl2 protein from an integrated data source STRING (Snel et al. 2000) produced a protein - protein network The yielded result from STRING is then imported into Cytoscape (Shannon et al. 2003). And, this network is further enhanced or reconstructed as in **Figure 49** using Cytoscape integrating information from different as well as related sources that allows the united network construction.

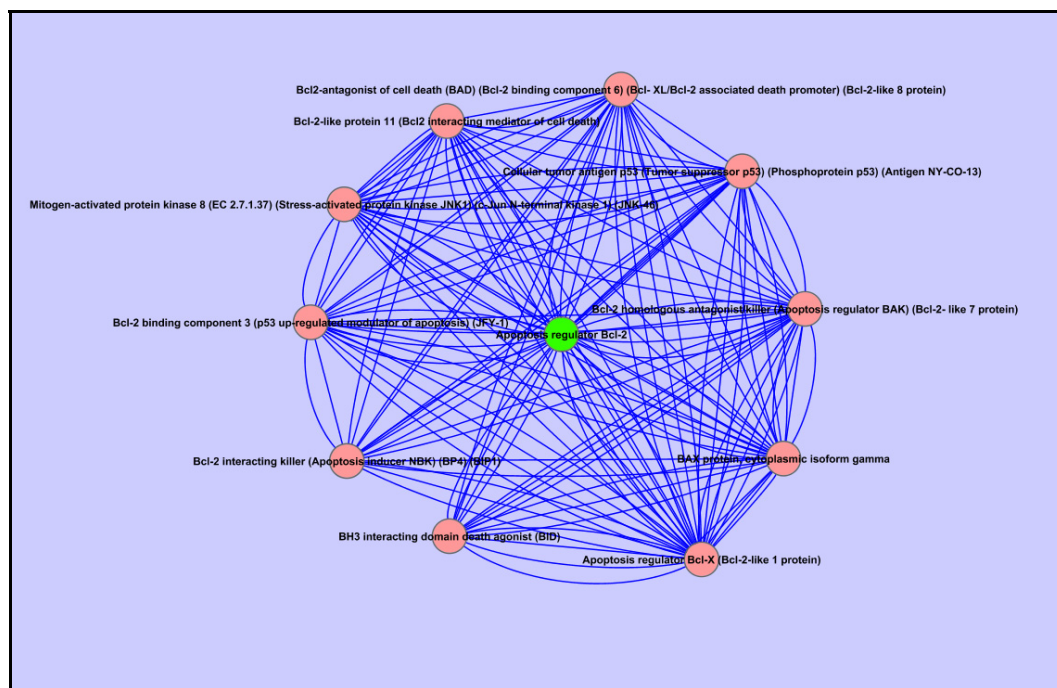


Figure 49: Protein - protein interaction network from an integrated source for Bcl2 along with MAPK and P53.

This unified network is a result of fusion methods (**Figure 50**) and it is coupled with protein-protein interaction networks generated from STRING consisting of several proteins that can give a global view of the role of the protein associated with different diseases and pathways along with more significant information such as its behaviour. Further by integrating protein-protein interaction network with the results or the related data from VINEdb, MoVisPP it paves the way for a greater or large scale network construction.

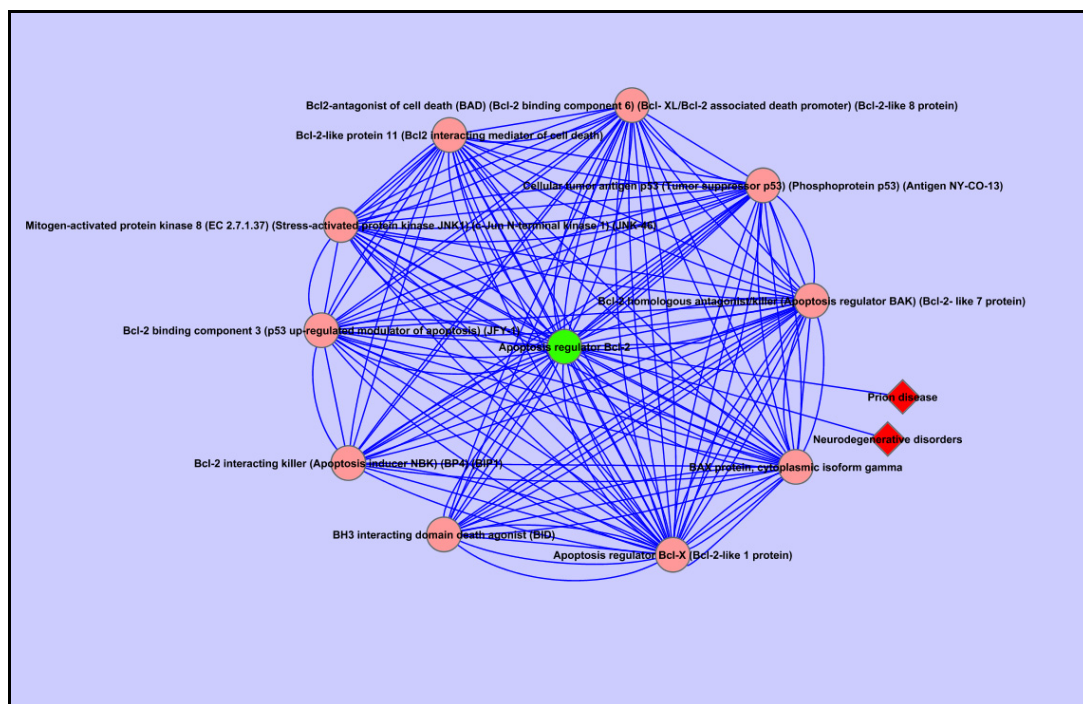


Figure 50: An enhanced protein – protein interaction network showing the associated diseases for Bcl2.

The construction and reconstruction of large scale biological networks using an integrative approach is further demonstrated by **Figure 51** as an example. This network is a combination of the integrative approach using a combination of semi-automatic, automatic and manual approach. The method follows the unification of multifarious biological data that are available in different databases which store the unique information about the proteins, genes, diseases pathways, ontology etc. for building the biological networks along with encapsulated information. Here the protein-protein interaction data generated from STRING is combined with other

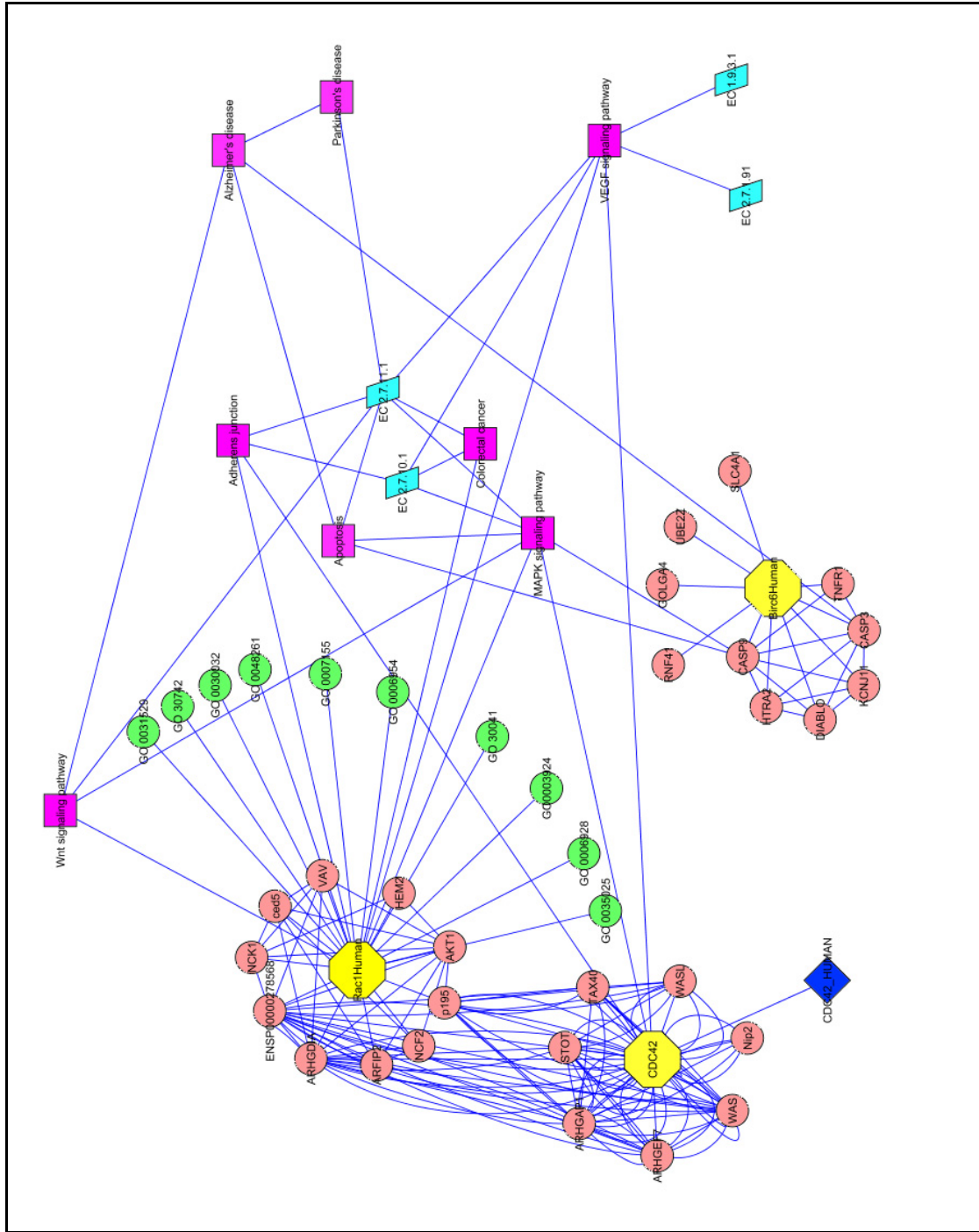


Figure 51: An expanded integrated network showing the protein-protein interaction networks for Human Rac1, Cdc42 and Birc6. information pertaining to the genes, protein, pathway, diseases, enzyme, ontology etc. from various other sources including VINEdb, MoVisPP.

The large scale network is constructed based on the proteins that participate in cardiovascular and other diseases. As shown the protein in the protein-protein interaction network shares some features / function with apoptosis, MAPK, cancer

Nature Precedings : doi:10.1038/npre.2010.5164.1 : Posted 2 Nov 2010

pathways. This network is constructed manually by using Cytoscape by the fusing the protein-protein interaction data along with other information. Network is composed of proteins and their interaction partners along with their Gene Ontology processes and related pathways. Further these pathways have been connected to enzymes to which they interact. Some of the diseases and pathways mentioned is also related and share some features to the above mentioned enzymes and proteins possibly based on their function or other characteristics.

The proteins Rac1, Cdc42 and Birc6 share common proteins which interact with each other. They are also the key players in cardiovascular diseases according to the experts have different neighbours and they are also closely or distantly associated with the diseases Alzheimer's, Parkinson's and Colorectal cancer which has been in the networks. Also, they are associated with apoptosis and other signaling pathways MAPK, VEGF etc. which play a key role in myriad of diseases along with other enzymes. MicroRNAs are a class of small non-coding RNAs (ncRNAs). They induce translation repression or mRNA degradation by binding to mRNA. It is also found that miRNAs could cause cancers acting as oncogenes or tumor suppressor genes. ncRNAppi (Ng et al. 2009) is a web-based disease-related miRNA target pathway database which links the miRNA target genes with their PPI partners based on the specific tissue or disease. A search for the genes Rac1, Birc6, Cdc42 and Bcl2 in ncRNAppi lists some of the associated and also common disease pathways. Following **Table 9** shows the genes and their associated pathways from ncRNAppi.

Table 9: Table lists the genes and miRNA targeted disease pathways

Gene	Related Disease Pathways
Bcl2	hsa01510, hsa04210, hsa04510, hsa05030, hsa05060, hsa05210, hsa05215, hsa05222
Birc6	hsa04120
Cdc42	hsa04010, hsa04360, hsa04370, hsa04510, hsa04520, hsa04530, hsa04660, hsa04670, hsa04810, hsa04912, hsa05120, hsa05130, hsa05131, hsa05211, hsa05212
Rac1	hsa01510, hsa04210, hsa04510, hsa05030, hsa05060, hsa05210, hsa05215, hsa05222

As discussed in the previous paragraphs and also from other studies it is established that apoptosis and MAPK have role in cardiovascular disease. Referring to the works of Berger and Iyengar and it is proposed by this approach and the network when enhanced at a major level with more information and analysis can help in designing

the drugs for cardiovascular diseases. With more understanding of the finer details and the nuances of the cardiac and other diseases and the pharmacological aspects it is possible to design drugs for cardiovascular and other related disease using a network biology approach.

The network throws light in a new direction i.e. towards an integrative approach for the construction and reconstruction of biological networks at large scale. This also provides a platform to understand the behaviour of the proteins when they are alone as well as when they work in unison with many other proteins and it is a part of the complex biological system or process. It is not possible to generate the quantitative (kinetic) parameters for all the proteins, reactions every time. Most of the parameters are still unknown or yet to be deciphered in order to built a biological network with all the quantitative data or information. The generated network is a qualitative network which does not need or depend on any quantitative information. Moreover, this network can be simulated if converted with appropriate tools and provided with the kinetic parameters that govern their reactions and relations. Further this network can be expanded at a larger scale using automated methods and integrating more data sources can help in the research studies. In a nutshell this network is the result of the “*Integration of in Silico and in Vitro Analyses*” and it paves the way for yet another dimension in drug design.

6.2 Summary

The large scale construction / reconstruction of biological network is yet another approach or steps towards the modelling and simulation in systems biology in order to understand the behaviour of the proteins when they are alone and work in concert with other proteins. This approach using combinational methods (manual, semi-automatic and automatic) helps to construct the biological process in various contexts integrating different information from various sources. The networks constructed and demonstrated in the earlier sections are the result of the fusion methods to bring in the protein-protein interaction, ontology, enzyme, gene, protein, pathways and disease data and how they share this information with each other in to perform a task. Moreover by providing the quantitative parameters and using appropriate tools it is possible and can be demonstrated that modelling and simulation of such biological networks in large scale is possible. Also it is possible the unified network

reconstruction of biological network using an integrative approach demonstrated can be approached and demonstrated at a more large scale for the construction of global biological maps.

It is demonstrated by combining both the *in Silico and in Vitro Analyses* i.e. the experimental data and the results generated from the tools (VINEdb, MoVisPP and other tools) it is possible to put forth a hypotheses which can help in the drug design using the network biology approach. This network produced gives multifarious information about the participating proteins at the same time along with their behaviour. This can help the biologists and other experts in the project, *CardioWorkBench - Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analyses* to view a global picture of the medically important proteins (Bcl2, Rac1, Cdc42 and many other proteins) and their behaviour independently and in concert with other proteins and in complexes. By this approach it will be helpful to know the various activities of the protein at a single given instance which can help in enhancing the knowledge which can be applied further to the design processes in this and other similar projects.

Chapter 7

Hypotheses and Discussion

The rudimentary biological data is generated by different experiments and the results are published in different formats and are also available in scientific literatures. The data from these sources are further curated and annotated by experts and then brought together by manual and automated methods through data integration approaches, further giving rise to several databases and data warehouses. It can be called data storage houses as they store petabytes and terabytes of data from the different sources. These mega warehouses can serve several applications and smaller databases and can exchange its data. At the same time the data integrated by these methods should have a goal and a purpose so that it can serve the community to a greater extent. In general it should be a concept driven approach that can bring in the right information together and avoid the unessential data that may be available and are not required for the task.

Furthermore these databases or data warehouses store the data in different formats most of them store in a tabulated form. These formats can be understood only by the creator of those databases or those who know computer science. In other way it can be accessed and understood only by the experts who work with these databases and understand and know how to query and store the data using a special language. One should be well aware of these standards and languages if they would like to work with these data storage information systems. It cannot be expected a scientist or an end user will know about the query languages and ways to access the data. They would like to work on these systems to know about the data and to use them further for their research. So it is not possible for them to access the data as such in its special format.

For this the tables of biological data that is housed in data warehouses can be integrated and further presented in a very user comprehensible manner. This makes things very easy to the user who can access the data through common web browser but does not want to know the backend changes and query language talks that takes place while they access and would like to get their desired results. When these biological databases apart from having the integrated data from multiple sources if possess a unique feature, can have edge or added advantage. If the feature is user

friendly then it will make the database or data warehouse or the application more accessible.

Alignment is a method to find out evolutionary relationship among proteins and pathways or to decipher the homology. By developing algorithms or concepts for the alignment it is possible to decipher the evolutionary relationship among the proteins, pathways and their phylogeny. Many works have been done in this direction to know their relationship and it is one of those well known and well researched areas in bioinformatics / computational biology. But this alignment approach if combined with a method to project the results in a very friendly manner where the end user could understand the relationship of pathways with ease can be very advantageous. The same alignment when combined with information from different sources can give more understanding of the relationship of the proteins participating in those biological pathways and about their homology.

Modelling and simulation of biological pathways has grown and it is one of the popular areas in the recent past where it has gained attention from different areas and it is undergoing an interdisciplinary approach. Experts from different areas try to build biological models using their domain knowledge and further combining with biological aspects they bring in more solutions. Many tools have been developed that can model and simulate the biological pathways using different parameters e.g. kinetics, enzyme concentrations etc. And it is not possible to derive the biological parameters for each molecule or entity that is involved in a pathway. These are labor intensive moreover sometime it is very hard to estimate the kinetics of the reactions etc. Also to construct a network from the scratch is itself an art as it needs expertise and knowledge of the domains. At the end of building a particular network it would have consumed more time and labor. The network whether in small scale or large scale is quite complex. So if the concept or hypothesis that can model and simulate biological pathways with ease is always welcome among the community as it involves less time and labor.

Based on these ideas several applications that are developed have been discussed in the earlier chapters. All these works try to focus on the concepts that lie beneath them and at the same time not losing the focus that it must be user comprehensible.

As a result Chapter 3 deals with a data warehouse that allows network navigation and interactive exploration of the integrated data by the user. The data warehouse combines the information from different source like other databases and it has been fortified with a visualization and monitor components that can project the integrated data in an effective manner and also take care of the information about the data source. Chapter 4 discusses and addresses the concept of alignment of biological pathways using protein structural information. The tool projects the result based on relational coloring pattern that is the result of the underlying protein structural information and classification scheme. More data sources are also combined that can substantiate the concept and evolutionary relationship among the pathways. Chapter 5 addresses the issue of large scale modelling of biological networks that is not dependent on the parameters discussed rather it is able to construct qualitative networks that will allow further modeling and simulation of the biological networks using Petri net based software suites. Chapter 6 deals with an application case to address a combinational or integrative approach for the construction / reconstruction of biological networks at a larger scale using the multifarious biological data to help in the drug design for the cardiovascular diseases and similar projects.

In a nutshell the construction and reconstruction of biological networks based on integrated methods has an added advantage compared to the conventional methods which consumes manual labor and needs biochemical parameters. Using integrated methods combining with visualization and different concepts provided a new dimension for building and extending biological networks. Moreover it is made possible to prove that visualization of diverse molecular data as a composite network image can have greater impact and understanding compared to the tabular form of representation of the data available in databases. Further the construction of biological networks without the need of biochemical parameters and at the same time fusing them with related data gives more understanding of the context and relevance of each participating molecule, relation, reaction etc. All these networks have many features in common that usually helps to compare them across species. And comparing them based on the protein structural information can have benefits not provided by other information. Furthermore combining all these features it was still made possible to make a network fusion approach where it further proves the *whole is more than the sum of its parts*.

Chapter 8

Conclusions

Everyday biological experiments lead to hundreds of scientific publications. At the same time it is very hard to understand each and every literature for scientists who work in these domains. Furthermore to develop tools and concepts based on these outputs or results is another milestone. It is always good to bear in mind that the applications developed must be user friendly so that the end users can work and understand the systems with ease. This also makes these applications access more by the users. Further care must be taken when the concept that underlie should be present in a simple way to the community. In the previous chapters several tools and databases were discussed. All the tools are user comprehensible at the same time they are being developed based on different concepts. Even though they differ in their concepts they share common features i.e. visualization and data integration approach. The work is a multi-pronged approach towards the analysis and representation of biological data. Also, by implementing the tools it is possible to answer those questions discussed in the earlier chapters and in the problem section (1.2).

(a) How to integrate the multifarious biological data and what are the methods that can be adapted?

By deploying the data warehouse in VINEdb it was not only possible to integrate the various data sources but also to monitor the source upgrades / updates. This helps to keep the warehouse abreast of the changes that happen in the source data. And the deployment of individual databases for SignAlign and MoVisPP integrating diverse information from different sources make it possible to Align and extend the KEGG based pathways and also bring in the related information.

(b) How to present the complex molecular data in an effective and comprehensible manner to the users?

All the tools/applications (VINEdb, MoVisPP, SignAlign) along with application case described before, use diverse data from various sources. They have their own logic and algorithm to handle and project the results. Yet, they are powered with a different visualization technique which allows them, to analyze and project their results in a user comprehensible manner. Instead of having complex tables or browsing several

pages these applications were able to give their results in an effective and user friendly way.

(c) Utilizing the diverse data how they can be transformed into biological networks and perform modelling and simulation of the pathways?

The composite networks of VINEdb gives a better understanding of the related molecular data but it is not possible to handle the static network with other software suites. By implementing MoVisPP which is fortified with the methods of data integration, visualization along with the methods to model and simulate the biological pathways it has become possible. The tool utilizes the information from the underlying data warehouse and constructs the networks based on the integrated diverse data. The extended networks can be exported to other software suites like Cell Illustrator which will allow the incorporation of biochemical parameters and modelling of the networks.

(d) How to tackle the problem of non-availability of quantitative data and still construct networks effectively and project the essence of complex nature of the data?

MoVisPP does not require quantitative data, at present, to construct the biological networks. Instead it constructs large scale biological networks using qualitative methods. It is able to generate network based on KEGG maps along with other related information. Thus, it is possible to maintain the complex nature and also the essence of data.

(e) What are the methods and information that can be helpful to know the evolutionary relationship among proteins and pathways?

Several works have been done which are discussed in the previous chapters to align or to detect the homology between the pathways. They have been using different data and approaches such as sequence and E.C. number based information. Some of the tools use various methods to detect the evolutionary relationship among pathways.

(f) How much is the protein structural information useful in deciphering the homology among biological pathway?

The protein structural information can be of significant value. Sometimes structure has more valuable information than sequence. Even with less sequence identity structural similarity is more. The protein structural information is utilized in SignAlign for the alignment of biological pathways. It is possible to decipher the

homologous pathways using this approach and it is based on the protein structural classification information.

Furthermore the following sub-sections summarize the works of different tools and data warehouse which are implemented.

Chapter 3 Relational data integration visualization and network navigation and exploration

The chapter deals with the data warehouse that is developed by bringing in data from several external data sources such as KEGG, OMIM, Intact, GO, UNIPROT and more. The tool VINEdb discussed show how the integrated data sources can be presented to the end user in a comprehensible manner. It is fortified with a visualization component that allows the user to navigate and explore the networks and their underlying data sources with ease instead of presenting the same data in a tabulated form. Furthermore the system is fortified with a monitor component that helps to know the up-to-dateness of the data from the external sources. With this approach a method to integrate several data sources and a method to explore the network with ease is demonstrated. An example to show the ability of the application is detailed. Further the data from this warehouse along with other sources is used and extended into a network. This further demonstrates the construction and reconstruction of biological networks using an integrative approach.

Chapter 4 Structure based Information and Integration for the Alignment of Biological Pathways

The concept of alignment is very historical and it is a method to find the homologous relationship between proteins and then the concept is extended to know the pathway evolution. Proteins share certain common features when they are evolutionarily related to each other. This is termed “homology”. Homology can be demonstrated at the sequence, structure and functional level. Many concepts and algorithms have been developed to understand the homology of proteins and pathways. Some works focused to understand pathway homology using Enzyme Classification. The chapter dealt with the alignment method for finding the evolutionary relationship between the pathways using protein structural information. The tool SignAlign is developed based on the concept and it demonstrates the method of aligning biological pathways using the underlying integrated data from different sources such as SCOP, CATH, PDB,

GO, ENZYME. Further it is connected to several external sources i.e. PROCOGNATE and QSCOP that can provide additional information about the protein that participates in the pathway. By this approach yet another direction is explored that can throw light towards the evolutionary relationship of biological pathways.

Chapter 5 Petri net based reconstruction and visualization of biological pathways using integrated molecular data

Modelling and simulation of biological networks has a significant role to understand the ability and biological function and complexity. Many tools and databases have been developed. Petri net is a mathematical representation of biological complexity that is easy to understand for biologists. Petri based tools follows certain basic features to model and simulate the networks and it is manually intensive to construct these networks. The chapter dealt with the large scale modelling and simulation of biological networks using integrated molecular data. MoVisPP, Modelling and Visualization of Pathways using Petri nets is based on the concept of data integration methods combined with construction of Petri net based KEGG pathways on the fly. This tool is very user friendly as it involves few simple steps to construct the Petri net based networks. Further it provides a window for the modelling and simulation of large scale biological networks as it allows the export of the constructed pathways to other Petri net based simulation tools such as Cell illustrator that allows further interaction with the constructed pathways.

Chapter 6 Construction and reconstruction of biological networks – an integrative approach

The larger scale construction and reconstruction of biological networks demonstrated as an application is an integrative approach. Using this method it is possible to explain how the proteins behave with each other and how much they can share their features with their neighbours or counterparts. The networks are a result of a fusion of the multifarious biological data that are generated from different sources to project an abstract image of the complexity of the biological system or the processes. For this network, data from VINEdb, MoVisPP, STRING, and other sources are integrated. The protein / gene around which the network is constructed are involved in cardiovascular and other diseases. This has been proved by various approaches.

Further this network biology approach can be helpful to the experts who are participating in the EU Project: *CardioWorkBench - Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analyses* and in other similar projects.

Chapter 9

Perspectives

Works demonstrated in the previous chapters show the ability of data integration and visualization methods combined with the concepts. The applications have been developed based on different algorithms and have been developed for different purposes i.e. construction and reconstruction of biological networks, alignment of biological pathways and network fusion. The work is a multi-pronged approach towards the analysis and representation of biological data. These works can be further extended and can lead to different directions and hypotheses in future. Like

- A common platform can be developed to integrate the different approaches.
- Trying to find a consensus across the different levels of protein comparison i.e. sequence, structure, enzyme and function, to find out their evolutionary relationship.
- Large scale integration and construction and reconstruction of biological pathways using multifarious data and approach.
- New methods of visualization and concepts can be developed.
- Interdisciplinary approach for the modelling and simulation of the networks for further understanding of biological function.

Glossary

Algorithm – Is a sequence of actions which perform a particular task.

Alignment – Is the adjustment or the arrangement of an object in relation with other objects or can also be defined as a orientation of an object or set of objects with respect to other objects.

Amino acid – are alpha amino substituted carboxylic acids which are the building blocks of proteins.

Apoptosis – programmed cell death in which a cell responding to a signal from outside or also programmed in its own genes, brings its own death and lysis by degrading its macromolecules systematically.

Architecture – Defines the style and method of construction of a system.

Cardiovascular – Is the circulatory system which comprises of the heart and blood vessels that carries nutrients and oxygen to the various tissues of the body and also removes carbon dioxide and other wastes from them.

Cascade – is a series of reactions involving regulatory events where one enzyme activates another which in turn activates a third enzyme which often happens by phosphorylation.

Caspases – Is a family of proteases which bring about programmed cell death.

CATH – Class Architecture Topology Homology is a database providing the information about proteins.

Cell cycle – It is sequence of events within the cell occurring between cell divisions.

Cellular differentiation – is a process of acquiring a new complement and RNA in which a precursor cell to attain a particular function becomes specialized.

Client – A computer or the software which runs on a computer and also interacts with another computer (server) located a distant or remote site.

Data warehouse - Is a repository of significant data of an organisation or project.

Database – Is a collection of data on a specific topic stored in an organized manner.

Disease – It is a pathological condition of a organ or system of an organism that is the results of various causes such as infection, genetic defect or environmental stress and other factors and further characterized by an identifiable or pronounced group of signs or symptoms.

Domain – Is a compact structural subunit of a protein.

Drug - Any substance that alters the normal bodily function when absorbed into the body of an organism.

Drug design – Is an inventive process that involve finding new medications based on the acquired knowledge of the biological target.

Drug target – It is a protein the function of which will be useful to alleviate or to modify a symptom of a disease.

Enzyme – is a biomolecule which can be a protein or RNA which is able to catalyze a chemical reaction. But it would not affect the equilibrium of the chemical reaction rather it enhances the reaction rate by furnishing a reaction path with a activation energy at a lower level.

Evolution – Is the change that occurs from one generation to the next in the genetic material in the population of organism

G protein – Are a family of cell signalling proteins which are regulated by guanine nucleotide binding.

Gene – Is a segment of chromosome that codes for a RNA molecule or single functional polypeptide chain.

Gene expression – It is transcription and in the case of proteins it is translation to yield a gene product; when its biological product is present and active then a gene is expressed.

Genotype – The genetic composition of an organism.

Graph – Is an abstract structure which consists of nodes and edges.

Homologous – Descended from a common ancestor

Homologous proteins – Are proteins having similar sequences and function in various species.

Insilico - Performed on computer or via computer.

Invitro – “in life” i.e. in the living cell or organism.

Invivo - “in glass” – that is inside the test tube.

Kinetics – Is study of reaction rates.

Ligand – Is a small molecule that is specific and binds only to the large molecule.

Macromolecule – Are those molecules which have few thousand to several millions of molecular weight range.

MAP Kinases – Are the ubiquitous regulators of cell growth and differentiation and is a family of mitogen-activated protein-serine/threonine kinases.

Metabolism - Is the sum of anabolism and catabolism and is the complete set of enzyme-catalyzed transformation of organic molecules in living cells.

Metabolite – In metabolism it is the intermediate chemical in the enzyme-catalyzed reactions.

Metabolome – Under a given specific conditions in a given cell it is the complete set of small molecule metabolites.

Motif – Is a pattern or finger print or the recurrence of an element.

Multiple Sequence Alignment – It is the assignment of residue correspondence mutually in more than a pair of sequences.

Negative feedback - A reaction product inhibits an earlier step in the pathway and it is a regulation of biochemical pathway.

NMR – It is a technique in order to study the structure and dynamics of the molecules which utilizes the quantum mechanistic properties of atomic nuclei of which they are a part.

Nucleic acids – These are biologically occurring polynucleotides in which nucleotide residues are connected in a certain sequence by phosphodi-ester bonds, DNA and RNA.

Pairwise sequence alignment – In two related sequences it is the assignment of residue correspondences.

Path – In a graph it is a consecutive set of edges.

Pathway – A sequence of biochemical or enzymatic reactions or events that converts one biological material to another.

Phylogeny – Refers to the evolutionary development of an organism.

Phenotype – Is the physical appearance of an organism.

Primary structure – It is the set of chemical bonds in a protein or nucleic acid.

Protein – Is a macromolecule which is composed of one or several polypeptide chains each of them having a characteristic sequence of amino acids that are linked by polypeptide bonds.

Protein Kinases – In a target protein these enzymes transfer the terminal phosphoryl group of ATP or another nucleoside triphosphate to a Ser, Thr, Tyr, Asp or His side chain leading to the regulation of the activity or other properties of that protein.

Protein-Protein interaction – Is the direct contact or association of protein molecules or sometimes through other medium for long range interactions.

Proteome - Is the complete complement of proteins that can be expressed by a given genome or in a given cell it is the full complement of proteins which is expressed.

Gene regulatory network – In a cell it is the collection of DNA segments which interact with each other and also with other substrates and governs the rate at which the genes in the network are transcribed into RNA.

RMSD – Root Mean Square Deviation is the measure of the differences between values predicted by a model and the actual values observed from the protein being modelled.

SCOP – Structural Classification of Proteins is a database containing the protein classification information.

Signalling transduction – Refers to the conversion process with which a cell converts any kind of stimulus or signal to another.

Signalling network – Are intertwined network which is formed by the interactions of several signalling pathways within a cell.

Similarity – It is the resemblance which is not necessarily obtained from homology.

Tertiary structure –Is the spatial arrangement of a polymer chain of a protein or nucleic acid.

Systems biology – Is a study of complex biochemical systems which integrates the functions of several to all of the macromolecules in a cell.

X-ray Crystallography - To determine the arrangement of individual atoms within a molecule it a method which uses the diffraction pattern of X-rays.

Bibliography

- Abyzov A and Ilyin VA. (2007) A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Structural Biology* 7:78.
- Ahouse JC. (2002) Gene regulation and metabolism. Edited by Collado-Vides J and Hofestädt R. *MIT press pp 2-5*.
- Almaas E. Biological impacts and context of network theory (2007). *J Exp Biol. May;210(Pt 9):1548-58*.
- Altermann E, Klaenhammer TR. (2005) PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics. May 3;6(1):60*.
- Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res. 2008 Jan;36(Database issue):D419-25*
- Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res. Jan 1;32(Database issue):D115-9*.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.(2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet. May;25(1):25-9*.
- Aytuna AS, Gursoy A, Keskin O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics. Jun 15;21(12):2850-5*.
- Bader GD, Cary MP, Sander C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res. Jan 1;34(Database issue):D504-6*.
- Bairoch A. (2000) The ENZYME database in 2000. *Nucleic Acids Res. 28:304-305*.
- Baitaluk M, Qian X, Godbole S, Raval A, Ray A, Gupta A.(2006) PathSys: integrating molecular interactions graphs for systems biology *BMC Bioinformatics, 7: 55*.
- Bakheet TM, Doig AJ. (2009) Properties and identification of Human Protein Drug Targets. *Bioinformatics Feb 15;25(4):451-7*.
- Bashton M, Nobeli I, Thornton JM. (2006) Cognate ligand domain mapping for enzymes. *J Mol Biol. Dec 8;364(4):836-52*.
- Bashton M, Nobeli I, Thornton JM.(2008) PROCOGNATE: a cognate ligand domain mapping for enzymes. *Nucleic Acids Res. 2008 Jan;36.(Database issue):D618-22*.
- Berger SI, Iyengar R. (2009) Network analyses in systems pharmacology. *Bioinformatics. Oct 1;25(19):2466-72*.

- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. *Nucleic Acids Res. Jan 1*;28(1):235-42.
- Birkland A, Yona G. (2006) BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics 7*:70, 2006.
- Blake JA, Bult CJ. (2006) Beyond the data deluge: Data integration and bio-ontologies. *J Biomed Inform. Jun*;39(3):314-20.
- Boulton SJ, Vincent S, Vidal M. (2001). Use of protein-interaction maps to formulate biological questions. *Curr Opin Chem Biol. Feb*;5(1):57-62.
- Boyadjiev SA, Jabs EW. (2000) Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin Genet. Apr*;57(4):253-66.
- Bray D (1995) Protein molecules as computational elements in living cells. *Nature. 1995 Jul 27*;376(6538):307-12.
- Brazhnik O, Jones JF. (2007) Anatomy of data integration. *J Biomed Inform. Jun*;40(3):252-69.
- Bruggeman FJ, Westerhoff HV. (2007) The nature of systems biology. *Trends Microbiol. Jan*;15(1):45-50.
- Calder M, Duguid A, Gilmore S, Hillston J. (2006) Stronger Computational Modelling of Signalling Pathways Using Both Continuous and Discrete-State Methods. In *CMSB (Priami, C. Ed.), volume 4210 of Lecture Notes in Computer Science, Springer, 2006.*
- Cao SL, Qin L, He WZ, Zhong Y, Zhu YY, Li YX. (2004) Semantic Search among Heterogeneous Biological Databases Based on Gene Ontology. *Acta Biochim Biophys Sin (Shanghai). 2004 May*;36(5):365-70.
- Cases I, Pisano DG, Andres E, Carro, A, Fernández, JM, Gómez-López, G, Rodriguez JM, Vera JF, Valencia A, Rojas AM. (2007) CARGO: a web portal to integrate customized biological information. *Nucleic Acids Res. Jul 1*;35 (Web Server issue):W16 - 20.
- Chaouiya C. (2007) Petri net modelling of biological networks. *Brief Bioinform. 2007 Jul*;8(4):210-9.
- Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. (2007) MINT: the Molecular INTERaction database. *Nucleic Acids Res. Jan*;35(Database issue):D572-4.
- Chen M and Hofestadt R. (2004) PathAligner: Metabolic Pathway Retrieval and Alignment. *Appl. Bioinformatics, 3*(4), 241-52.
- Chen M, Hariharaputran S, Hofestadt R, Kormeier B and Spangardt S. (2009) Petri Net Models for the Semi-automatic Construction of Large Scale Biological Networks. *Nat Comput DOI 10.1007/s11047-009-9151-y.*
- Chi PH, Pang B, Korkin D, Shyu CR. (2009) Efficient SCOP-fold classification and retrieval using index-based protein substructure alignments. *Bioinformatics. Oct 1*;25(19):2559-65.
- Cho SG, Choi EJ. (2002) Apoptotic signaling pathways: caspases and stress-activated protein kinases. *J Biochem Mol Biol. Jan 31*;35(1):24-7.

- Choi C, Münch R, Leupold S, Klein J, Siegel I, Thielen B, Benkert B, Kucklick M, Schobert M, Barthelmes J, Ebeling C, Haddad I, Scheer M, Grote A, Hiller K, Bunk B, Schreiber K, Retter I, Schomburg D, Jahn D. (2007) SYSTOMONAS--an integrated database for systems biology analysis of Pseudomonas. *Nucleic Acids Res.* 2007 Jan;35(Database issue):D533-7.
- Choi K and Kim S.(2008). Compath: comparative enzyme analysis and annotation in pathway/subsystem contexts. *BMC bioinformatics* 9, 145.
- Chou CH, Chang WC, Chiu CM, Huang CC, Huang HD. (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res.* 2009 Jul 1;37(Web Server issue):W129-34.
- Collado-Vides J and Hofestädt R (2002). Gene regulation and metabolism. *MIT Press.*
- Collado-Vides J, Hofestädt R, Mavrovouniotis M, Michal G.(1999) Modeling and simulation of gene regulation and metabolic pathways. *Biosystems.* Jan;49(1):79-82.
- Coulombe B, Blanchette M, Jeronimo C. (2008) Steps towards a repertoire of comprehensive maps of human protein interaction networks: the Human Proteothèque Initiative (HuPI). *Biochem Cell Biol.* 2008 Apr;86(2):149-56.
- Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA. (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* Jan;37(Database issue):D310-4.
- Dandekar T, Schuster S, Snel B, Huynen M, Bork P (1999). Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J. Oct 1;343 Pt 1:115-24.*
- de la Fuente A, Fotia G, Maggio F, Mancosu G and Pieroni E (2008). Insights into biological information processing: structural and dynamical analysis of a human protein signalling network. *J. Phys. A: Math. Theor.* 41 224013.
- Dobson PD and Doig AJ (2005) Predicting enzyme class from protein structure without alignments. *J Mol Biol.* Jan 7;345(1):187-99.
- Eidhammer I, Jonassen I and Taylor WR (2004) Protein bioinformatics. *John Wiley & Sons, Ltd.*
- Elmasri R and Navathe SB (2000). Fundamentals of database systems. *Addison Wesley, 843 – 844.*
- Fietz C. (2007) Ein Data-Warehouse-Ansatz zur Integration und Visualisierung von biomedizinischen Daten. *Diplomarbeit, AG Bioinformatik, Technische Fakultät, Universität Bielefeld.*
- Fischer M, Thai QK, Grieb M, Pleiss J. (2006) DWARF – a data warehouse system for analyzing protein families. *BMC Bioinformatics, 7: 495.*
- Funahashi A, Tanimura N, Morohashi M, and Kitano H. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks, *BIOSILICO, 1:159-162.*
- Gansner ER and North SC (1999) An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper., 00(S1), 1–5.*

- George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB. SCOPEC: a database of protein catalytic domains. *Bioinformatics*, 20(1), 130-136, 2004.
- Getz G, Starovolsky A, Domany E. (2004) F2CS: FSSP to CATH and SCOP prediction server. *Bioinformatics*. Sep 1;20(13):2150-2.
- Ghazal P. (2008) Pathway Biology Approach to Medicine. *Handbook of Research on Systems Biology Applications in Medicine* By Andriani Daskalaki.
- Golemis E. (2002) Protein-Protein interactions edited by Erica Golemis, *CSHL Press, 2002*.
- Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, Hollmén J, Oresic M. (2005) Data integration and visualization system for enabling conceptual biology. *Bioinformatics Jun;21 Suppl 1:i177 - 85*.
- Goss PJ, Peccoud J. (1998) Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc. Nat. Acad. Sci. Vol. 95, 6750 – 6755*.
- Grafahrend-Belau E, Schreiber F, Heiner M, Sackmann A, Junker BH, Grunwald S, Speer A, Winder K, Koch I. (2008) Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics. Feb 8;9:90*.
- Green ML, Karp PD. (2006) The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Res. Aug 7;34(13):3687-97*.
- Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC. (2001) DiscoveryLink: A System for Integrated Access to Life Science Data Sources. *IBM Systems Journal, 40(2):489-511*.
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. (2005) Online Inheritance in Man (OMIM), a knowledgebase of human gene and genetic disorders. *Nucleic Acid Res. 33: D514-D517*.
- Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. (2007) The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol. Apr;8(4):319-30*.
- Hanahan D, Weinberg RA. (2000) The hallmarks of cancer. *Cell. Jan 7;100(1):57-70*.
- Hardy S and Robillard PN. (2004) Modeling and Simulation of Molecular Biology systems using Petri nets: Modeling Goals of Various Approaches. *J Bioinform Comput Biol. 2004 Dec;2(4):595-613*.
- Hardy S and Robillard PN. (2007). Visualization of the simulation data of biochemical network models: a painted Petri net approach. *SCSC 2007*.
- Hariharaputran S, Töpel T, Oberwahrenbrock T and Hofestädt R (2007) SignAlign: Prediction and alignment of biochemical pathways using protein structural information. *In proceedings of the 5th International Conference on Pathways, Networks, and Systems, Porto Heli, Greece*.
- Hariharaputran S, Töpel T, Oberwahrenbrock T and Hofestädt R. (2008) Alignment of Linear Biochemical Pathways Using Protein Structural Classification. *Nature Precedings, doi:10.1038/npre.2008.1943.1*.

- Hariharaputran S, Töpel T, Brockschmidt B and Hofestädt R. (2007) VINEdb: a datawarehouse for integration and interactive exploration of life science data. *Journal of Integrative Bioinformatics* 4(3):63.
- Hirsh E, Sharan R (2007) Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics* Jan 15;23(2):e170-6.
- Ho J W E, Manwaring T, Hong S H, Roehm U, Fung D C Y, Xu K, Kraska T, Hart D. (2006) PathBank: Web-Based Querying and Visualziation of an Integrated Biological Pathway Database. *cgiv*, pp.84-89, *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*.
- Hofestädt R and Scholz U. (1998) Information processing for the analysis of metabolic pathways and inborn errors. *Biosystems* 47, 91 - 102.
- Hofestädt R and Thelen S. (1998) Quantitative modeling of biochemical networks. *In Silico Biol.* 1(1):39-53.
- Holm L, Kääriäinen S, Rosenström P, Schenkel A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*. Dec 1;24(23):2780-1.
- Holm L, Sander C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* Jan 1;25(1):231-4.
- Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U. (2006) COPASI--a COMplex PATHway SIMulator. *Bioinformatics*. Dec 15;22(24):3067-74.
- Hopkins AL. (2007) Network pharmacology. *Nat Biotechnol.* Oct;25(10):1110-1.
- Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* Jul 1;37(Web Server issue):W115-21.
- Hu Z, Mellor J, Wu J and DeLisi C. (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5, 17.
- Huang H, Barker WC, Chen Y et al. (2003) iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.* 31(1): 390-392.
- Iragne F, Nikolski M, Mathieu B, Auber D and Sherman D. (2005) ProViz: protein interaction visualization and Exploration. *Bioinformatics* Jan 15;21(2):272 4.
- Janes KA, Yaffe MB. (2006) Data-driven modelling of signal-transduction networks. *Nat Rev Mol Cell Biol.* Nov;7(11):820-8.
- Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, Liu B, Nandi A, Santos C, Andrews P, Athey B, States D, Jagadish HV. (2007) Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Res.* Jan; 35 Database issue: D566-71.
- Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* Jan;37(Database issue):D412-6.
- Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L.

- (2005) Reactome: a knowledge database of biological pathways. *Nucleic Acids Res. Jan 1;33(Database issue):D428-32.*
- Ju BH, Han K. (2003) Complexity management in visualizing protein interaction networks. *Bioinformatics. 2003;19 Suppl 1:i177-9.*
 - Junker BH, Klukas C, Schreiber F. (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics. Mar 6;7:109.*
 - Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res. Jan;36(Database issue):D480-4.*
 - Kanehisa M, Goto S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res. Jan 1;28(1):27-30.*
 - Karp P (1995) A strategy for database interoperation. *Journal of Computational Biology, 2(4):573-586.*
 - Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W83-8.*
 - Khan TA, Bianchi C, Ruel M, Voisine P, Sellke FW. (2004) Mitogen-activated protein kinase pathways and cardiac surgery. *J Thorac Cardiovasc Surg. Mar;127(3):806-11.*
 - Kim WK, Bolser DM, Park JH. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics. 2004 May 1;20(7):1138-50.*
 - Koonin EV, Wolf YI. (2006) Evolutionary systems biology: links between gene evolution and function. *Curr Opin Biotechnol. 2006 Oct;17(5):481-7.*
 - Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res. Jan 1;34(Database issue):D689-91.*
 - Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol. Dec;23(12):1509-15.*
 - Lee DY, Saha R, Yusufi FN, Park W, Karimi IA. (2009) Web-based applications for building, managing and analysing kinetic models of biological systems. *Brief Bioinform. Jan;10(1):65-74.*
 - Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics. Mar 23;7:170.*
 - Lenzerini M (2002). Data Integration: A Theoretical Perspective. *PODS 2002. pp. 233-246.*
 - Lesk AM. (2005) Introduction to Bioinformatics. *Oxford University Press.*

- Li W, Liu Y, Huang H, Peng Y, Lin Y, Ng W and Ong K (2007). Dynamical systems for discovering protein complexes and functional modules from biological networks. *IEEE-ACM transactions on computational biology and bioinformatics*, vol. 4, no. 2, pp. 233-250.
- Liang Z, Xu M, Teng M and Niu L. (2006) NetAlgin: A web-based tool for comparison of protein interaction networks. *Bioinformatics*, 22(17), 2175-77.
- Lo Conte L, Brenner SE, Hubbard TJ, Chothia C and Murzin A. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, 30(1), 264-7.
- Maurizio L. (2002) Data Integration: A Theoretical Perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 233 – 246.
- Mayo M (2005). Learning Petri net models of non-linear gene interactions. *Biosystems* 82, 74-82.
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res. Jul;35(Web Server issue):W182-5*.
- Murata T. (1989) Petri Nets: Properties, Analysis and Applications. *In: Proceedings of the IEEE, Vol. 77, No. 4, pages 541-580*.
- Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
- Muslin AJ. (2008) MAPK signalling in cardiovascular health and disease: molecular mechanisms and therapeutic targets. *Clin Sci (Lond)*. Oct;115(7):203-18.
- Nagasaki M, Doi A, Matsuno H and Miyano S. (2004) Integrating Biopathway Databases for Large-scale Modeling and Simulation. *In Proc. Second Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand. CRPIT, 29. Chen, Y.-P. P., Ed. ACS. 43-52*.
- Nagasaki M, Doi A, Matsuno H, Miyano S. (2005) Petri net based description and modeling of biological pathways. *Algebraic Biology - Computer Algebra in Biology*, pp. 19 – 35.
- Ng A, Bursteinas B, Gao Q, Mollison E, Zvelebil M. (2006) Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Brief Bioinform.* Dec;7(4):318-30.
- Ng KL, Liu HC, Lee SC.(2009) ncRNAppi--a tool for identifying disease-related miRNA and siRNA targeting pathways. *Bioinformatics.* 2009 Dec 1;25(23):3199-201.
- Noble D. (2003) The future: putting Humpty-Dumpty together again. *Biochem Soc Trans.* Feb;31(Pt 1):156-8.
- Oehm S, Gilbert D, Tauch A, Stoye J, Goesmann A. (2008) Comparative Pathway Analyzer--a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms. *Nucleic Acids Res.* Jul 1;36(Web Server issue):W433-7.

- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. (1997) CATH--a hierarchic classification of protein domain structures. *Structure*. Aug 15;5(8):1093-108.
- Pavlopoulos GA, Gap, Wegener AL, Aw, Schneider R. (2008). A survey of visualization of tools for biological network analysis. *BioData Min*. Nov 28;1(1):12.
- Pieroni E, de la Fuente van Bentem S, Mancosu G, Capobianco E, Hirt H, de la Fuente A (2008). Protein networking: insights into global functional organization of proteomes. *Proteomics*. 2008 Feb;8(4):799-816.
- Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M. (2005) Alignment of metabolic pathways. *Bioinformatics*. Aug 15;21(16):3401-8.
- Pireddu L, Poulin B, Szafron D, Lu P, Wishart DS (2005) Pathway Analyst — Automated Metabolic Pathway Prediction. In *Proc. of the IEEE 2005 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*.
- Pireddu L, Szafron D, Lu P, Greiner R. (2006) The Path-A metabolic pathway prediction web server. *Nucleic Acids Res*. Jul 1;34.
- Qi Y, Ge H. (2006) Modularity and dynamics of cellular networks. *PLoS Comput Biol* 2(12), 2006.
- Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*. 2005 Apr 1;21(7):1189-93.
- Reddy VN, Mavrovouniotis ML, Liebman MN. (1993) Petri net representations in metabolic pathways. *Proc Int Conf Intell Syst Mol Biol*. 1:328-36.
- Rice JJ, Stolovitzky G. (2004) Making the most of it: pathway reconstruction and integrative simulation using the data at hand. *Drug Discovery Today: BIOSILICO*. Vol 2, Issue 2, Pages 70-77.
- Rocha J, Segura J, Wilson RC, Dasgupta S. (2009) Flexible structural protein alignment by a sequence of local transformations. *Bioinformatics*. Jul 1;25(13):1625-31.
- Runge T. (2004) Application of Coloured Petri Nets in Systems Biology. *Proc. 5th Workshop CPN, Univ. of Aarhus, October 2004*, pp. 77–95.
- Sackmann A, Heiner M, Koch I. (2006) Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*. Nov 2;7:482.
- Saraiya P, North C, Duca K. (2005) Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Information Visualization*, Vol. 4, No. 3. 191-205.
- Schadt EE, Lum PY. (2006) Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res*. Dec;47(12):2601-13.
- Shah PK, Patrick Aloy P, Peer Bork P and Robert B. Russell RB. (2005) Structural similarity to bridge sequence space: Finding new families on the bridges. *Protein Science*, 14:1305-1314.

- Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. (2005) Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*. Feb 21;6:34.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. Nov;13(11):2498-504.
- Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. (2005) Conserved patterns of protein interaction in multiple species. *PNAS*. 102, 6,1974-1979.
- Sierk ML, Pearson WR. (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci*. Mar;13(3):773-85.
- Sirava M, Schäfer T, Eiglsperger M, Kaufmann M, Kohlbacher O, Bornberg-Bauer E, Lenhof HP. (2002) BioMiner--modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*.;18 Suppl 2:S219-30.
- Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US. (2003) The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics*. 2003 Feb 12;19(3):408-15.
- Snel B, Lehmann G, Bork P, Huynen MA. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*. Sep 15;28(18):3442-4.
- Song N, Joseph JM, Davis GB, Durand D. (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol*. May 16;4(4):e1000063.
- Spangardt S. (2007) Ein webbasiertes Tool zur Modellierung und Visualisierung biochemischer Pathways als Petri-Netze. *Diplomarbeit, AG Bioinformatik, Technische Fakultät, Universität Bielefeld*.
- Stebbings LA, Mizuguchi K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res*. Jan 1;32(Database issue):D203-7.
- Stein LD. (2003) Integrating biological databases. *Nat Rev Genet*. May;4(5):337-45.
- Stein M, Gabdoulline RR, Wade RC. (2008) Calculating enzyme kinetic parameters from protein structures. *Biochem Soc Trans*. 2008 Feb;36(Pt 1):51-4.
- Strange K. (2005) The end of "naive reductionism": rise of systems biology or renaissance of physiology?. *Am J Physiol Cell Physiol*. May;288(5):C968-74.
- Stumpf MP, Kelly WP, Thorne T, Wiuf C. (2007) Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol Evol*. Jul;22(7):366-73.
- Suhrer SJ, Wiederstein M, Sippl MJ (2007) QSCOP - SCOP Quantified by Structural Relationships. *Bioinformatics* 23(4):513-514.
- Teichmann SA, Grishin NV. (2006) Sequences and topology: from methods to meaning. *Curr. Op. Struc. Biol.*, 16, 359-261.

- Trissl S, Rother K, Müller H, Steinke T, Koch I, Preissner R, Frömmel C, Leser U. (2005) Columba: an integrated database of proteins, structures, and annotations. *BMC Bioinformatics*. 2005 Mar 31;6:81.
- Uetz P, Finley RL Jr. (2005) From protein networks to biological systems. *FEBS Lett*. Mar 21;579(8):1821-7.
- Uetz P, T Ideker, B Schwikowski. (2002) Visualization and integration of protein-protein interactions. In: *E Golemis (ed.): Protein-Protein Interactions - A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press, 623-646.
- UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*. Jan; 37(Database issue):D169-74.
- van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C.(2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*. Sep 25;9:399.
- Voet D, Voet JG, Pratt CW. (1999) Fundamentals of Biochemistry. pp 354-355. *John Wiley & Sons*.
- Wang E, Lenferink A, O'Connor-McCourt M. (2007) Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol Life Sci*. 2007 Jul;64(14):1752-62.
- Wang Y. (2007) Mitogen-activated protein kinases in heart development and diseases. *Circulation Sep 18;116(12):1413-23*.
- Watterson S, Marshall S, Ghazal P. (2008) Logic models of pathway biology. *Drug Discov Today*. May;13(9-10):447-56.
- Wu J, Mao X, Cai T, Luo J, Wei L.(2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res*. 2006 Jul 1;34(Web Server issue):W720-4.
- Wu J, Voit E. (2009) Hybrid modeling in biochemical systems theory by means of functional petri nets. *J Bioinform Comput Biol*. Feb;7(1):107-34.
- Ye Y, Doak TG (2009) A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLoS Comput Biol* 5(8): e1000465. doi:10.1371/journal.pcbi.1000465.
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. (2002) MINT: a Molecular INTERaction database. *FEBS Lett*. Feb 20;513(1):135-40.

Appendix I

Abbreviation

CATH	Class Architecture Topology Homology
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
UniPROT	Universal Protein Resource

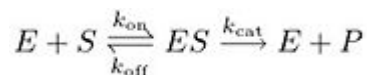
Appendix II

Websites

CATH	http://www.cathdb.info/
GO	http://www.geneontology.org/
IntAct	http://www.ebi.ac.uk/intact/main.xhtml
KEGG	http://www.rcsb.org/pdb/home/home.do
MoVisPP	http://agbi.techfak.uni-bielefeld.de/movispp/
OMIM	http://www.ncbi.nlm.nih.gov/omim/
PDB	http://www.rcsb.org/pdb/home/home.do
PROCOGNATE	http://www.ebi.ac.uk/thornton-srv/databases/procognate/
QSCOP	http://qscop.services.came.sbg.ac.at/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
SignAlign	http://agbi.techfak.uni-bielefeld.de/signalign/index.jsp
UniPROT	http://www.uniprot.org/
VINEdb	http://agbi.techfak.uni-bielefeld.de/VINEdb/

Appendix III

As explained in the SBML web pages the following enzymatic reaction



will be represented by SBML as

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml level="2" version="3"
xmlns="http://www.sbml.org/sbml/level2/version3">
  <model name="EnzymaticReaction">
    <listOfUnitDefinitions>
      <unitDefinition id="per_second">
        <listOfUnits>
          <unit kind="second" exponent="-1"/>
        </listOfUnits>
      </unitDefinition>
      <unitDefinition id="litre_per_mole_per_second">
        <listOfUnits>
          <unit kind="mole" exponent="-1"/>
          <unit kind="litre" exponent="1"/>
          <unit kind="second" exponent="-1"/>
        </listOfUnits>
      </unitDefinition>
    </listOfUnitDefinitions>
    <listOfCompartments>
      <compartment id="cytosol" size="1e-14"/>
    </listOfCompartments>
    <listOfSpecies>
      <species compartment="cytosol" id="ES" initialAmount="0"
name="ES"/>
      <species compartment="cytosol" id="P" initialAmount="0"
name="P"/>
      <species compartment="cytosol" id="S" initialAmount="1e-
20" name="S"/>
      <species compartment="cytosol" id="E" initialAmount="5e-
21" name="E"/>
    </listOfSpecies>
    <listOfReactions>
      <reaction id="veq">
        <listOfReactants>
          <speciesReference species="E"/>
          <speciesReference species="S"/>
        </listOfReactants>
        <listOfProducts>
          <speciesReference species="ES"/>
        </listOfProducts>
        <kineticLaw>
          <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
              <times/>
              <ci>cytosol</ci>
              <apply>
                <minus/>
                <apply>
                  <times/>
```

```

        <ci>kon</ci>
        <ci>E</ci>
        <ci>S</ci>
    </apply>
    <apply>
        <times/>
        <ci>koff</ci>
        <ci>ES</ci>
    </apply>
</apply>
</math>
<listOfParameters>
    <parameter id="kon" value="1000000"
units="litre_per_mole_per_second"/>
    <parameter id="koff" value="0.2"
units="per_second"/>
</listOfParameters>
</kineticLaw>
</reaction>
<reaction id="vcat" reversible="false">
    <listOfReactants>
        <speciesReference species="ES"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="E"/>
        <speciesReference species="P"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
                <times/>
                <ci>cytosol</ci>
                <ci>kcat</ci>
                <ci>ES</ci>
            </apply>
        </math>
        <listOfParameters>
            <parameter id="kcat" value="0.1"
units="per_second"/>
        </listOfParameters>
    </kineticLaw>
</reaction>
</listOfReactions>
</model>
</sbml>

```

Short Vita

January 2005 to present

PhD Student, member of the Bioinformatics / Medical Informatics Department and Graduate College Bioinformatics (GK635) at Bielefeld University, Germany.
Advisor - Prof. Dr. Ralf Hofestädt

Prior to joining the PhD program gained research experience working with *Prof. Nagasuma R Chandra* at Supercomputer Education and Research Centre / Bioinformatics Centre at the Indian Institute of Science, Bangalore and earlier in the Computational Neuroscience Laboratory working with *Prof. Upinder S Bhalla* at National Centre for Biological Sciences, Bangalore, India after obtaining the Master of Science (M.Sc.) degree from the University of Madras, Chennai, India followed by a Post Graduate Diploma in Computer Applications.

Awards and Fellowships

- The Newton International Fellowship for post-doctoral research awarded jointly by the UK's national research academies - The British Academy, The Royal Academy of Engineering and The Royal Society, 2009. (Selected)
- Fellowship to participate in the 5th International Conference on Pathways, Networks and Systems, 24 - 29 June 2007, Porto Heli, Greece.
- From January 2005 to December 2007 - PhD Scholarship, funded by DFG Graduiertenkolleg Bioinformatik (GK635), Bielefeld University, Germany.
- Awarded fellowship to attend The Ninth Workshop on Software Platforms for Systems Biology held in October 2004, Heidelberg, Germany.