WWW GUIDE

# Protein structure prediction on the Web

*Lorenzo Segovia*

Determining a protein's three-dimensional structure is an extremely time-consuming process, but the wealth of information that is obtained is well worth the effort, as it opens up new avenues of research through which to understand structure–function relationships. Unfortunately, the tools for such structural analyses—X-ray crystallography and NMR—are beyond the reach of many research groups, but in their absence, an alternative exists: structure prediction. Although the models obtained are only predictions, they very often give good working insights into the protein in question.

Until very recently, making secondary structure predictions using multiple sequence alignments or running fold recognition algorithms was only feasible with powerful workstations. With the advent of the WWW, however, anyone with a personal computer connected to the Internet can now have access to this kind of computer power through friendly and comprehensive interfaces. In this way, "wet bench" scientists can obtain a series of predictions that allow them to view their favorite proteins from a three-dimensional perspective instead of the one-dimensional view of the amino acid sequence. This permits interrelation of distant sequence regions that are actually close in the protein structure, thus possibly explaining some obscure mutant phenotypes. Even a sketchy knowledge of the spatial distribution of amino acid residues permits the design of rational and efficient directed-mutagenesis approaches to a desired phenotype.

These approaches are not exclusive to biochemically characterized proteins; indeed, fold recognition can help to assign a function to a hitherto poorly understood enzyme. Such analyses have now been undertaken by a variety of groups to characterize full genomes. For the uninitiated, several courses on structure prediction have been published on the WWW. These allow a researcher to follow a simple pathway from primary structure through secondary structure prediction all the way to a tertiary structure prediction.

A general strategy might be the following: The initial step is searching in databases for homologues to the test sequence. These searches lead to the identification of domains or sequence motifs (e.g. nucleotide binding), and the construction of multiple alignments of the homologs. If a homolog is found in a protein database (PDB), it is quite easy to obtain a three-dimensional picture by homology modeling. If there are no discernible homologs, a more complicated course needs to be followed: First, one must obtain a secondary structure prediction, and then a protein fold recognition (or threading) must be performed. If a particular fold gets a good score, the test sequence's predicted secondary structure elements are aligned against the target's and then the test sequence is aligned against the structure. A three-dimensional protein model is then obtained by comparative modeling.

*Lorenzo Segovia, department of molecular recognition and biostructure, Instituto de Biotecnologia, UNAM, Mexico (lorenzo@ibt.unam.mx).*

## A sampler of products and service sites related to protein structure prediction

| | | |
|---|---|---|
| **Introductory course** | | |
| A guide to structure prediction | http://bonsai.lif.icnet.uk/people/rob/CCP11BBS/ | A very concise yet comprehensive introductory course for protein structure prediction. |
| **Domain or motif identification** | | |
| The ProDom protein domain | http://protein.toulouse.inra.fr/prodom.html | A searchable protein domain database constructed by an all against all search of the Swissprot database. |
| Multiple EM for Motif | http://www.sdsc.edu/MEME/meme/website/ | MEME is a tool for discovering motifs in a group of elicitation (MEME) related protein sequences. These motifs can then be used to search against sequence databases and to construct multiple sequence alignments. |
| Blocks WWW Server segments | http://www.blocks.fhcrc.org/ | Blocks is a database of multiply aligned ungapped corresponding to the most highly conserved regions of proteins originally based on the Prosite database. |
| **Secondary structure prediction** | | |
| The PredictProtein server (PHD) | http://www.embl-heidelberg.de/predictprotein/phd_pred.html | PredictProtein computes a multiple sequence alignment and predictions of secondary structure, residue solvent accessibility, and location of transmembrane helices. |
| Self-optimized prediction method e-mail server | http://www.ibcp.fr/serv_pred.html | This server calculates a consensus secondary structure prediction based on five different methods. |
| **Fold recognition or tertiary structure prediction** | | |
| UCLA-DOE structure prediction server | http://www.mbi.ucla.edu/people/frsvr/frsvr.html | Fold recognition based on sequence-derived predictions. |
| TOPITS | http://www.embl-heidelberg.de/predictprotein/phd_pred.html | Fold recognition by prediction-based threading (one of the options in the PHD server). |
| Threading 123-D | http://www-lmmb.ncifcrf.gov/~nicka/123D.html | 123D uses a substitution matrix, secondary structure prediction, and contact capacity potentials to thread a sequence through the set of structures. |