GENOMICS

Bring on the promoters

High-throughput saturation mutagenesis determines the contribution of each base in a core promoter to overall promoter strength.

The more genomes are sequenced, the clearer it becomes that variation, particularly in regulatory regions, abounds. This raises the question of what the functional impact of mutations in these regions truly is. As Jay Shendure from the University of Washington put it, "It has been challenging to tie this variation back to function in an efficient way... we need methods to analyze function." Heeding this call, his team devised a high-throughput method to screen for the effect of mutations in core promoters.

Their goal was to test the effect of all possible single-nucleotide mutations in a core promoter in a single experiment. The researchers synthesized the mutant promoters—three bacteriophage and three mammalian core promoters—on programmable microarrays that yield 200-base oligonucleotides. Each promoter was flanked by its native sequence and a unique barcode identifying the mutation. After releasing the oligonucleotides from the arrays, the researchers transcribed them *in vitro* and sequenced the cDNA on Illumina's Genome Analyzer. The abundance of each barcode provided a digital readout for how well each promoter supported transcription.

The team chose well-characterized promoters and sought to validate their results against what has been reported in the literature. Many of their findings correlated with expectations; for example, mutations in the TATA box sharply reduced transcription. Some were surprising, such as the observation that some double mutants were more robust and showed higher activity than either of the single mutants. The researchers hypothesized that whereas one mutation increases the binding of the polymerase, thereby decreasing transcriptional activity, a second mutation might weaken the binding again, thus counteracting the effect of the first.

One current constraint of the method is the length of the oligonucleotides that can be programmed on the array. "A functional promoter goes well beyond the core region," says Shendure, and he acknowledges that the array-based synthesis technique his team uses is currently not amenable to longer oligonucleotides. As an alternative, he is considering assembly in solution, which would allow the building of longer constructs.

One direction the researchers want to pursue is assaying the functionality of common variation in regulatory sequences, specifically enhancers and untranslated regions affecting transcript stability, an area that Shendure describes as understudied. He recognizes that longer oligonucleotides are not the only improvement needed: "To do all of this convincingly," he says, "we at least have to move into cell lines." Indeed, transfecting hundreds of barcoded oligonucleotides into cells, rather than just transcribing them *in vitro* with a cell extract, and then retrieving the transcripts for high-throughput sequencing will require some improvement to the method.

But besides being a tool for large-scale functional analysis for naturally occurring

PROTEOMICS

THE IMPORTANCE OF BEING NEGATIVE

The Negatome is a database of non-interacting protein pairs that can be used for training protein-protein interaction prediction algorithms.

Who cares about negative results? It's fairly safe to say that most researchers would not try to publish a paper that focused on what they did not find, and that even if they did try, they would be hard-pressed to find a journal that would agree to publish it. However, that is not to say that negative results do not have scientific value—in fact, they can be quite useful.

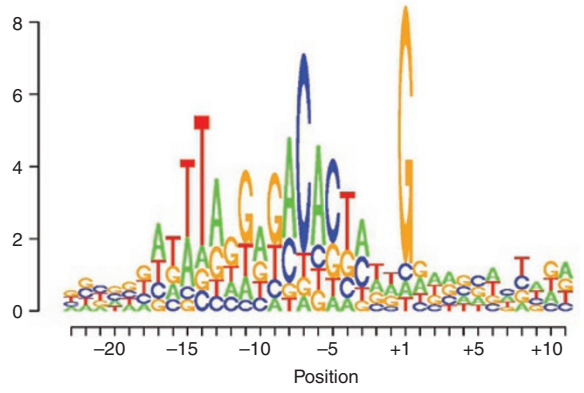
In the field of bioinformatics, for example, both a positive and a negative dataset are required for training machine learning algorithms, such as protein-protein interaction prediction tools. Dmitri Frishman, along with co–first authors Pawel Smialowski and Philipp Pagel and their colleagues at the Technical University of Munich and the German Research Center for Environmental Health, recently introduced a well-curated database of protein pairs that are unlikely to interact, aptly called the "Negatome."

Protein interactions are responsible for carrying out almost all biological functions; the entire network of interactions is known as the interactome. We are still very far away from mapping the entire interactome of any cellular organism, so good prediction tools to generate hypotheses are needed. But although well-curated positive datasets of protein-protein interactions exist for several organisms, defining with certainty which proteins do not interact is actually extremely difficult. In addition to the slim literature evidence, "there is no technique that can conduct a large-scale measurement of non-interacting pairs," explains Frishman.

In the past, others have used information about cell localization to construct a negative training dataset, based on the hypothesis that proteins found in different cellular compartments are unlikely to interact. But this is not ideal, explains Frishman: "If you use protein localization to train your predictor, you end up with a predictor of co-localization and not necessarily of interaction between proteins."

Frishman and his colleagues thus constructed the Negatome using two different types of information. One, they combed the literature for negative interaction data from individual experiments. Not surprisingly, they did not find a large number of negative results, but using stringent criteria for selection, they generated a list of 1,162 negative interaction pairs. Two, they looked at structural data from the Protein Data Bank (PDB) to identify proteins that participated in the same protein complex but did not directly interact; this list comprised 745 protein pairs. Only 15 protein pairs from the literature-curated and PDB-curated datasets overlapped, for a total of 1,892 non-interacting protein pairs.

When they compared their non-interacting protein pairs to the STRING database, a vast resource containing both experimental and predicted protein-protein interactions based on physical,



Activity logo for a particular promoter. Reprinted from *Nature Biotechnology*.

variations, Shendure also foresees that saturation mutagenesis will benefit the field of synthetic biology. "Groups that focus on synthetic circuits," he says, "are largely borrowing from nature's tool box. Less has been done to engineer them from the ground up, fine-tune, optimize or modulate those existing parts." This type of mutagenesis strategy would allow engineers to predetermine promoters with very specific properties, build an oligonucleotide library and characterize which mutants fit the requirements.

Genome biology and synthetic biology can share the same tool kit, one to understand nature's variation, the other to mimic and improve upon it.

Nicole Rusk

RESEARCH PAPERS

Patwardhan, R.P. *et al*. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).

genetic and functional evidence, the researchers observed that less than 10% of their negative pairs were functionally associated by STRING. However, because STRING is not just limited to physical interactions, functional associations are likely to yield false positive hits. The Negatome also certainly includes false "negative" information; it is surely possible that some of the negative interactions can indeed occur under some biological context.

In addition to training protein-protein interaction prediction algorithms, the Negatome could also be used to judge the quality of high-throughput interactome screens such as two-hybrid methods, which have been criticized for being subject to a high false positive rate. "If you think about these famous 'hairballs,' these huge networks of interactions, use of the Negatome would be a way to erase some of the edges, if a particular edge is stated as being false," notes Frishman.

The Negatome currently contains data mostly for mammalian proteins, but Frishman and his colleagues have longer-term plans to continue adding new literature evidence and structure-based data from the PDB, which will continually improve the resource. Perhaps the larger scientific community will see the value of the Negatome and thus be encouraged to make negative results, in many different fields, more widely available.

Allison Doerr

RESEARCH PAPERS

Smialowski, P. *et al.* The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res.* published online 17 November 2009.

NEWS IN BRIEF

GENOMICS

Complete genomes

Despite advances in high-throughput sequencing, the number of completely sequenced human genomes is still small. A new technique and business model by the company Complete Genomics promises to change that. Their technology involves the assembly of fragmented DNA into nanoballs that are arrayed and sequenced using combinatorial probe-anchor ligation chemistry. The low cost of consumables and efficient parallelization allow Complete Genomics to project that their company will sequence hundreds of individuals in the near future.

Drmanac, R. *et al. Science* advance online publication 5 November 2009.

PROTEOMICS

The human protein-DNA interactome

Although the DNA targets of key transcription factors have been intensively studied, the targets of the broader set of DNA-binding proteins are largely unknown. Hu *et al.* used a bioinformatics approach to predict human proteins likely to interact with a set of 460 diverse DNA motifs. They then used a protein microarray containing the 4,191 known and predicted DNA binding proteins to characterize the human protein-DNA 'interactome'; they identified a large number of known and previously unknown protein-DNA interactions.

Hu, S. *et al. Cell* **139**, 610–622 (2009).

BIOINFORMATICS

Correcting gene function annotations

Homology-based methods to annotate gene function are subject to misannotations that can propagate through databases; thus, they are very important to correct. Hsiao *et al.* describe an algorithm for policing gene annotations. The algorithm looks for genes with poor genomic correlations with their network neighbors, which are likely to represent errors. Hsiao *et al.* applied their approach to identify misannotations in *Bacillus subtilis*.

Hsiao, T.-L. *et al. Nat. Chem. Biol.* **6**, 34–40 (2010).

STEM CELLS

The fate of stem cells

There is high interest in understanding what happens, on a systems level, to stem cells upon perturbation. Lu *et al.* follow changes in histone acetylation, chromatin-bound RNA polymerase II, mRNA and nuclear protein levels in mouse embryonic stem cells after downregulation of the pluripotency factor Nanog. They find that this single perturbation has widespread repercussions across the epigenetic, transcriptional and translational systems.

Lu, R. *et al. Nature* **462**, 358–362 (2009).

GENOMICS

Dancing in the rain

When looking for variation in the human genome, researchers are often interested in just a specific subsection. But how best to enrich for such a region? Scientists from the company RainDance Technologies present a microdroplet-based technology in which each target region is amplified in a singleplex reaction within the confines of a microdroplet. This allows efficient amplification of target regions with uniform coverage, high accuracy and reproducibility.

Tewhey, R. *et al. Nat. Biotechnol.* **27**, 1025–1031 (2009).