J. Biol. Journal of Biology

BioMed Central

Research news

Tracking evolution's footprints in the genome

Jonathan B Weitzman

Published: 23 June 2003

Journal of Biology 2003, **2**:9

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/2/2/9

© 2003 BioMed Central Ltd

The strategy of using 'phylogenetic footprinting' to find regulatory sites that are conserved between pairs of related complex genomes has led to the development of a suite of computational tools that succeed in finding functionally important transcription-factor-binding sequences.

Celebrating the latest completed genome sequencing project is all very well, but even before the champagne runs dry questions are asked about how best to use all the sequence information. The observation that less than 2% of the human genome sequence actually encodes proteins is a sobering issue for the 'post-genomic era'. And finding functionally relevant information within the non-coding sequence presents a formidable challenge, akin to tracking footprints in a dense forest. In this issue of the *Journal of Biology* [1], Boris Lenhard, Albin Sandelin, Wyeth Wasserman and colleagues describe a computational approach that will benefit all researchers keen to locate and explore the regulatory elements in their chosen genome (see 'The bottom line' box for a summary of their work).

Predicting binding sites

Understanding the principles that govern where and when genes are expressed is essential for deciphering how genome information is turned into the molecular and cellular phenomena that underlie the biology of complex organisms. Gene expression programs are determined through the recognition of specific promoter and enhancer sequences within the DNA by regulatory transcription-factor proteins. **Transcription-factor-binding sites** (**TFBSs**; see the 'Background' box) are short sequences, many of which have been painstakingly elucidated over the years using experimental procedures such as **DNAse footprinting** and **electrophoretic mobility shift assays (EMSA)**. TFBSs tend to be short, often less that 10 base-pairs long, and thus they are likely to occur within a genome quite often simply by chance. In addition, each transcription factor appears to tolerate a wide range of variations from its simple consensus sequence, making it extremely difficult

The bottom line

- Finding transcription-factor-binding sequences within DNA is difficult, because the sequences recognized by individual factors are short and not entirely conserved.
- Looking for potential transcription-factor-binding sites (TFBSs) that are conserved between two related genomes – 'phylogenetic footprinting' – improves predictions.
- The ConSite algorithm aligns non-coding orthologous sequences from two genomes and screens them against the JASPAR database, which comprises a library of experimentally verified TFBSs, to further improve the sensitivity and selectivity of predictions of TFBSs.
- The ConSite web interface allows all researchers to apply the algorithm to their genome(s) of interest, and to screen the database of experimentally verified TFBSs, providing useful tools for unraveling the mysteries of transcriptional regulation.

Background

- **Transcription-factor-binding sites (TFBSs)** are short sequences near the transcription-start site of each gene to which specific transcription-factor proteins bind.
- **DNAse footprinting** is an experimental technique used to identify the DNA region bound by a given transcription factor. It is often used with **electrophoretic mobility shift assays (EMSA)** to demonstrate specific DNA-protein interactions.
- A **position weight matrix (PWM)** is a statistical model that represents the frequency at which each nucleotide is observed at each position within a DNA sequence motif. These are used for computational prediction of putative TFBSs.
- **Phylogenetic footprinting** attempts to identify regulatory DNA sequences on the basis of their conservation in an alignment of genomic DNA from different species.
- ***Cis*-regulatory modules** are the clusters of TFBSs that regulate each gene, often including multiple sites for each transcription factor that regulates the gene.
- **ChIP-on-chip** or **ChIP-chip** is a recently developed technique that uses chromatin immunoprecipitation (ChIP) of transcription factors with their associated DNA, followed by microarray (DNA chip) analysis of the bound DNA sequences.

to predict binding sites by simply searching a genome sequence for consensus motifs.

“Characterization of the promoter regions of eukaryotic genes remains one of the most elusive problems in computational genome analysis,” says Roderic Guigó (Institut Municipal d’Investigació Mèdica, Barcelona, Spain). To address these challenges, bioinformaticians have developed approaches using **position weight matrices (PWMs)** that take into account the observed frequency of tolerated sequence variations at each nucleotide position within a consensus TFBS and give a quantitative score that reflects the actual binding specificity of the factor. Extensive investigation of transcriptional regulation has provided insights into how gene expression is finely regulated by the sequence and distribution of multiple TFBSs within *cis*-regulatory regions upstream of each gene. Combinations of TFBSs for different factors can form ***cis*-regulatory modules**, with complex functional synergy, that drive the transcriptional machinery.

The first thing that Wyeth Wasserman’s group did was build a library of high-quality PWMs. The quality of these matrices is critical for accurate site prediction. The best way to build a PWM is to plunge into the published literature and pull out relevant information from papers describing *in vitro* and *in vivo* experiments on individual transcription factors. “The collection of binding profiles, collectively termed the JASPAR database, was produced by the pure determination of Albin Sandelin for his thesis project studying the binding similarities of transcription factors in the same structural families,” says Wasserman. (See the ‘Behind the scenes’ box for further discussion of the motivation for the work.) The team constructed over a hundred binding-profile matrices for different transcription factors. Any DNA sequence can be screened using these matrices to locate potential TFBSs. A certain number of potential sites will be identified just by chance, however, and finding a potential site doesn’t guarantee that the cognate factor actually binds there or that the site is of biological relevance.

Two genomes are better than one

When the draft of the human genome sequence was published in 2001, David Baltimore wrote the following in an accompanying commentary [2]: “Gene-regulatory sequences are now there for all to see, but initial attempts to find them were also disappointing. This is where the genomic sequences of other species – in which the regulatory sequences, but not the functionally insignificant DNA, are likely to be much the same – will open up a cornucopia”. This is the basis of the method of **‘phylogenetic footprinting’**. The idea is that important regulatory modules are under selective pressure during evolution and that comparing two (or more) genomes will identify the conserved sequences that are most likely to be biologically relevant [3]. “Having multiple orthologous genes available provides a tremendous amount of information about what the most important features of the sequences are. It is the most valuable of ‘sequence only’ data,” says computational biologist Gary Stormo (Washington University School of Medicine, St Louis, USA). Guigó adds “in fact, we can say that without the genomes of other species, it will be impossible to fully understand the human genome.”

Having assembled the JASPAR database, the second feature of the Wasserman team’s approach was to create

Behind the scenes

Journal of Biology asked Wyeth Wasserman about how and why his group developed the ConSite suite of computational tools.

What motivated you to develop the ConSite system?
Originally there was a perception that the statistical models used to predict transcription-factor-binding sites were flawed, but several lines of evidence emerged to show that the models accurately reflected interactions outside cells. To improve predictions, my group developed several methods based on the study of combinations of binding sites for sets of transcription factors known to act together in specific types of cells. While these models are adequate, there are only a few tissue types with sufficient data to support their development. To overcome the specificity challenge for a broader range of researchers, we turned to phylogenetic footprinting.

How long did it take you to develop the system and what were the steps that ensured your success?
There were three critical components. First, we needed an alignment algorithm capable of accurately aligning long genomic sequences in reasonable time. Second, we required access to a collection of statistical models for a large set of transcription factors. Third, we needed a suite of bioinformatics methods to manipulate the alignments and models. Each of these was under development in the group prior to the conception of ConSite. In early 2001 we decided to combine the three pieces into a single system; by the summer we had it up and running. We waited a year for the compilation of the mouse genome sequence to provide the necessary data to quantitatively measure the performance of ConSite.

What were your initial reactions to the results and how has this approach been received by others in the field?
We knew where we were going, so there was no shock. But there is tremendous satisfaction to seeing everything come together, and this was amplified by the process. In bioinformatics, research success is often the result of a single person sitting in front of a computer. To make ConSite work, we had to work as a team. ConSite, TFBS and JASPAR have received outstanding support. The TFBS package is being used by researchers throughout the world. We are preparing to lead a tutorial on its use at upcoming bioinformatics conferences. The JASPAR database becomes available to the public with the publication of this article. We expect that it will also be used extensively.

What are the next steps and what does the future hold?
There are several key steps in the coming few years. First, the methods must be extended to handle the concurrent study of sequences from multiple species (instead of pairwise comparisons). Second, prediction of individual sites is still flawed and must be replaced by methods based on regulatory modules and clusters of transcription-factor-binding sites. Third, we need a larger database of binding profiles, which should emerge from the new 'ChIP-on-chip' studies. Finally, we must eventually develop a new generation of bioinformatics methods that address chromatin structure.

tools for aligning long stretches of genomic DNA. "The alignment algorithm by Luis Mendoza (originally called DPB and now re-engineered and named ORCA) is part of a bioinformatics system termed OrthoSeq that is undergoing final revisions," says Wasserman. Phylogenetic footprinting approaches have proved powerful in previous studies of particular genomic loci but have rarely been applied on a genome-wide scale [4-7].

The final challenge was to combine the genome-alignment tools with the PWMs to create a system that was easy to use. "The third component, the computer methods, were the focus of a project by Boris Lenhard to create a suite of computer programming resources for researchers engaged in the study of regulatory sequences. This system, the TFBS Perl module, has been available for about a year and is already being broadly used in the field," says Wasserman.

When these three elements were combined, ConSite was born [8]. The authors are eager for their tools to be widely used and have done their best to make them accessible and user-friendly. "This collection is a resource for the global bioinformatics community," says Wasserman. "As opposed to commercial databases of transcription-factor information, we make our data available without restriction to academic research groups. Consistent with the philosophy of *Journal of Biology* and the Public Library of Science [9], we believe in open data access."

Time for testing

With the ConSite suite of tools assembled, Lenhard *et al*. [1] conducted several tests to demonstrate the utility of their approach. They analyzed a number of well-characterized human gene promoter regions, comparing sequences with mouse and cow orthologs. They showed that adding the phylogenetic footprinting step improved the selectivity of TFBS prediction by 85% without a great loss of sensitivity. "Phylogenetic footprinting had already been postulated as a means to improve the characterization of the promoter regions of the genes in higher eukaryotic genomes, but the Wasserman article shows that the idea really works," says Guigó. Stormo comments that such programs cannot claim to be fully comprehensive; they will miss some sites, "but the sites that it does identify have a much greater probability of being important. So the reported sites will have a low false-positive rate, in contrast to some of the previous approaches".

The ConSite platform is likely to undergo many modifications and updates as bioinformaticians add new features and capabilities. The ability to align multiple sequences should further improve the phylogenetic footprinting selectivity. "[The authors] don't try to discover new types of sites, just to reliably identify the occurrences of sites for known transcription factors. But the approach can be extended to identifying new sites," says Stormo.

In the future, information from bioinformatic analyses might be combined with experimental datasets to construct models for complex transcriptional regulatory networks. Stormo envisages incorporating data from experiments using microarray analysis, **ChIP-on-chip** and mutant phenotyping to get a more complete picture of network connections. A recent study from Richard Young and colleagues [10] demonstrated how these approaches can be applied on a genome-wide scale in yeast.

Understanding the genetic networks regulated by transcription-factor activity will not only provide molecular insights into fundamental biological processes: it is also relevant to many disease pathologies and may perhaps indicate novel therapeutic strategies. Computational approaches such as ConSite will prove invaluable in this endeavor. Hunters of the past and present have always begun by tracking down the footprints. Now, genetic hunters have a powerful set of tools to help with their task.

References

1. Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis**. *J Biol* 2003, **2:**13.
2. Baltimore D: **Our genome unveiled.** *Nature* 2001, **409:**814-816.
3. Ureta-Vida AI, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes.** *Nat Rev Genet* 2003, **4:**251-262.
4. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
5. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-1394.
6. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
7. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003; Published online May 29; doi 10.1126/science.1084337
8. **ConSite** [http://www.phylofoot.org/consite]
9. **Public Library of Science** [http://www.plos.org]
10. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.

Jonathan B Weitzman is a scientist and science writer based in Paris, France.
E-mail: jonathanweitzman@hotmail.com