3D Protein Databases Integrated with Genomic and Chemical Data

D. Edwards, L. Yan, K. Olszewski and M. Donlan

Molecular Simulations Inc.

9685 Scranton Road, San Diego, CA 92103, USA

ABSTRACT

The function of a protein is determined more directly by its three-dimensional structure than its sequence, it is therefore particularly useful to know the structure of a protein. This has been demonstrated by the many successes of structural biology.

Initiatives are underway that use comparative protein structure modeling to generate structural data for all completely sequenced genomes through the use of an automated pipeline. Analysis of such vast amounts of structural data enables the development of a new type of 3D data that describes a given function. This data can then be used to predict the function of other novel protein targets and thereby provide a completely new approach to the identification of function. An alternative approach to using 3D information to identify function is through protein threading. An example of a protein threading algorithm which uses predicted secondary structure to identify the putative fold of a given protein sequence is given.

Integration of this type of three-dimensional protein information with genomic data can lead to a much clearer understanding of a protein's function and further focus the mining of chemical databases for structure-based drug discovery.

A description is given of the current state of homology-based data generation technologies. The developing techniques for characterization of active sites are discussed and evaluated in the light of the large volume of genomic data now available. How this data may be readily integrated into a drug discovery program is also highlighted.

Keywords: high throughput genome analysis, comparative modeling, database annotation.

INTRODUCTION

Functional genomics is the characterization of the proteins expressed by the genome. Because the function of a protein is determined more directly by its three-dimensional structure than its sequence, it is particularly useful to know the structure of these proteins. This has been demonstrated by the many successes of structural biology (e.g. Agouron's HIV protease inhibitor could not have been optimized in the absence of the receptor structure). Through in-house efforts and external collaborations MSI is developing methods which leverage structural information to assign function at a rate which is higher than the current methods used in traditional bioinformatics approaches. For example; SeqFold, a new application developed in conjunction with Professor David Eisenberg [1], uses predicted secondary structural information to assign the 3-dimensional fold of a protein and thereby detect remote homologies. Additionally, a new methodology that generates modeled structural data that can be used to aid the assignment and understanding of protein function at a genomic level has been developed by Andrej Šali and coworkers at Rockefeller University [2]. The method uses comparative protein structure modeling to generate structural data for a complete genome through the use of an automated pipeline. An increase of 30% in the number of sequence-structure relationships identified over standard sequence-searching methods was observed. These additional relationships include 3% of the genome which did not previously have a clear link to a protein sequence with known function. An additional 5% increase is anticipated from the incorporation of threading techniques [1] and hidden Markov models in the process of template identification. Analysis of such vast amounts of structural data would enable the development of structural templates that represent active site pharmacophores that describe a given function. These pharmacophores can then be used to predict the function of other novel protein targets and hence provide a completely new approach to the identification of function.

NEW METHODOLOGIES FOR ASSIGNING FUNCTION

Assigning function through fold recognition

SeqFold provides a new algorithm developed in the laboratory of Dr. David Eisenberg, which aids in the functional identification of proteins. One of the most promising approaches to fold recognition is based on the idea that sequence homology may be more sensitive and selective when aided by secondary structure information. Hence, the SeqFold sequence-structure similarity scoring function consists of two terms: sequence-based and structure-based similarity and the alignment optimization involves a sum of those terms. Obviously, any given novel sequence does not have a secondary structure annotation available therefore a secondary structure prediction algorithm such as GOR, DSC, or PHD, can be used to obtain approximate annotation. Fischer and Eisenberg have demonstrated that including secondary structure information improves detection of folds in a comprehensive fold recognition benchmark [3].

Additional validation studies have been carried out. One retrospective study demonstrates the knowledge gained over more commonly used sequence similarity searches. Leptin is a small hormone encoded by the *ob* gene; *ob*/*ob* mice are extremely obese and diabetic but when injected with leptin respond with rapid weight loss (hence the commercial interest in leptin). The leptin gene was identified via positional cloning in 1994 at Rockefeller University but, unfortunately, no clear sequence homology to any other protein with known structure was identified using standard bioinformatics techniques. However, SeqFold predicts that leptin belongs to the class of short-chain 4-helical cytokines. The structure of leptin, which has recently been solved by X-ray crystallography [4], reveals that the leptin is a long-chain 4 helical cytokine. The 3-dimensional model of leptin generated from the SeqFold sequence-structure alignment is observed to capture the essential features of the leptin structure such as buried aromatic residues and a disulfide bond that is necessary for function.

Since most of the commercially interesting genomic data is found in EST databases, a further study was performed in order to determine the usefulness of this application for determining function for protein sequences encoded by human EST data. A frequency analysis of the October 1998 release of gbEST revealed that between 70 and 80% of ESTs encode for a protein of sufficient length to represent a complete protein domain (80-100 amino acid residues). Since threading technologies, such as SeqFold, are most effective in recognizing relationships across full-length protein domains the results from this survey of gbEST confirm that this technology is appropriate for assigning function for translated EST data. Based on this conclusion, 41 ESTs were selected from gbEST that had no homology to any sequences of known function in SWISSPROT. SeqFold searches using the protein sequences translated from these ESTs identified clear homologies for 7 of the 41 sequences. The functional assignments ranged from cytokine to transcription regulation to toxin.

Assigning function by high throughput genome analysis

Andrej Šali and Roberto Sánchez at Rockefeller University have published a paper entitled "Comparative protein structure modeling in functional genomics" [2]. The paper details their efforts to identify and create homology models for as many Open Reading Frames (ORFs) as possible in a given genome. The initial study was performed using the *Saccharomyces cerevisiae* (Baker's yeast) genome for which it was possible to model substantial parts of 17.2% of the genome's ORFs. Although this percentage seems low, in actual fact 1,071 yeast protein models were constructed. Given that only 40 X-ray structures of yeast proteins are present in the Brookhaven Databank this represents a considerable increase in the structural information now available. In real terms, this represents an increase in the number of proteins for which a relationship with a known structure was assigned (over standard sequence searching methods) by 30% (236 proteins). The yeast genome is particularly well-characterized and therefore, an even greater increase in functionally identification is anticipated for other, less well-characterized genomes. The method used was able to establish clear functional assignments for an extra 3% of Baker's Yeast genome, beyond what had been previously identified in the literature. The process was repeated for four other genomes *E. coli*, *M. genitalium*, *C. elegans and M. janaschii*. The percentage of ORFs modeled ranged from 15.7% to 20.4%. The level of collation of existing scientific analyses for these four genomes does not enable similar estimates of what additional information was found using this method. In the study Šali used the PDB as a "virtual genome" test set and found that it was much more accurate to judge the validity of a sequence-structure match by evaluating the resultant homology models than by sequence similarity techniques alone. The number of false positives in this test study was less than 5%.

Using structural data in functional genomics

Comparative models are based on a sequence alignment between the protein to be modeled and a related protein of known structure. A question arises as to what additional insights that are not already possible from sequence matching alone can possibly be obtained by 3D modeling. The first advantage of 3D modeling is that it provides the best way of either confirming or rejecting a remote match, as described

in the previous section. This is important because most of the related protein pairs share less than 30% sequence identity. For example, only 10.7% of the yeast ORFs have been matched reliably with known structures by Fasta [5], as opposed to 17.2% in the study by Šali and Sánchez. Another case in point is that 236 of the 1071 yeast ORFs with predicted good models had no previously identified links to a protein of known structure in the major annotations of the yeast genome, including Sacch3D [6], Pedant [5], GeneQuiz [7], and PFAM[8]. Of these 236 proteins for which some structural information is now available, 41 also did not have a clear link to a protein sequence with known function.

The second advantage of 3D modeling over sequence matching is that some binding and active sites cannot possibly be found by searching for local sequence patterns [9,10], but frequently are detectable by searching for small 3D motifs that are known to bind or act on specific ligands [11]. This is a consequence of the facts that (i) structure is more conserved than sequence [12], (ii) 3D motifs tend to consist of residues distant in sequence, and (iii) there are some 3D motifs whose residues do not follow the same order in sequence, even though they have the same geometric arrangement. One example of this is the serine catalytic triad that almost certainly arose through convergent evolution in serine proteases of the trypsin and subtilisin type, and also in some lipases [11]. The 3D motifs could be defined in terms of features extracted from known protein-ligand structures, such as the constituting atoms and distances between them, shape, secondary structure, and electrostatic properties. Enumeration of active and binding sites for many proteins in the genome, such as various metal and nucleotide binding sites, will facilitate experimental determination of protein function.

The third advantage of 3D modeling over sequence matching is that a 3D model frequently allows a refinement of the functional prediction based on sequence alone because the ligand binding is most directly determined by the structure of the binding site rather than its sequence. An example of this is provided by a predicted SH3 domain in the yeast ORF[2]. Since there are known 3D structures of SH3 domains bound to proline-rich peptide ligands, it was possible to calculate a 3D model of such a complex for the putative yeast SH3 domain. Based on the model, the SH3 residues that interact with the peptide were identified. This result can be used to construct site-directed mutants with altered or destroyed binding capacity, which in turn could be used to test hypotheses about the sequence-structure-function relationships for this SH3 domain. In addition, such an analysis could increase or decrease the probability that a real protein-ligand pair has been found.

In the light of the fact that recognition of a molecule's fold itself may not be sufficient to assign function and that a protein's function is determined more directly by its three-dimensional structure than its sequence, we have developed methods to address this problem. First, a method of analysis of multiple families of sequences and associated structures using a method called Evolutionary Trace has shown that it is possible to recognize differences in functional specificity between sub-classes of proteins for a particular function that cannot be detected by sequence analysis methods alone [13]. Second, a tool which enables a user to generate a template based query that can be used to search for particular spatial arrangements of residues (e.g. the catalytic triad which characterizes a serine protease). Such pharmacophore-style representations can be further extended to identify glycosylation sites, for example, or phosphate, sulfate or metal binding sites. One can imagine being able to use the derived information to construct a query that relates other functional information with structurally annotated information. For example; "Show me all the proteins that are over-expressed in breast cancer, do not have a phosphate binding site (i.e. are not kinases) but that do have a metal binding site, or a site which I could engineer to bind a metal (given that metal binding sites can be used to turn certain types of activity on or off)."

Another advantage of a database of 3D protein structures is the ability to be able to perform a comparison of structures across a given protein family to highlight differences in the properties of binding site of a protein of interest (such as its shape, volume and electrostatics). These observations would enable a scientist to evaluate the potential of a protein as a drug target by evaluating the level of specificity for which a drug could be designed/hoped to bind. This assessment would be extremely useful in determining which protein targets from a given pathway or disease state should be prioritized as candidates for further screening or drug discovery.

REFERENCES

[1] D. Fischer and D. Eisenberg, P.N.A.S., 94, 11929-11934, 1997.

A. Šali and J.P. Overington, Protein Science, 3, 1582-1596, 1994.

[2] D. Fischer and D. Eisenberg, Protein Science, 5, 947-955, 1996.

[3] F.Zhang, M.B.Basinski, J.M.Beals, S.L.Briggs, L.M.Churgay, D.K.Clawson, R.D.Dimarchi, T.C.Furman, J.E.Hale, H.M.Hsiung, B.E.Schoner,

D.P.Smith, X.Y.Zhang, J.P.Wery,R.W.Schevitz Nature 387, 206, 1997.

[4] R. Sánchez and A. Šali, P.N.A.S. 95,13597-13602, 1998.

[5] D. Frishman and H. W. Mewes, PEDANT: Protein extraction, description, and analysis tool, URL: http://pedant.mips.biochem.mpg.de/frishman/pedant.html (1997).

[6] S. A. Chervitz, Sacch3D: Structural Information for Yeast Proteins, URL: http://genome-www.stanford.edu/Sacch3D (1997).

[7] M. Scharf *et. al*., GeneQuiz analysis of the S. cerevisiae genome, URL: http://www.sander.ebi.ac.uk/genequiz (1997).

[8] E. L. L. Sonnhammer, S. R. Eddy, R. Durbin, Proteins 28, 405 (1997).

[9] Bairoch, Nucl. Acids Res. 20, 2013 (1992).

[10] T. Pawson, Nature 373, 573 (1995).

[11] Wallace, N. Borkakoti, J. M. Thornton, Protein Sci. 6, 2308 (1997).

[12] Chothia and A. M. Lesk, EMBO J. 5, 823 (1986).

[13] O. Lichtarge, K.R. Yamamoto and F.E. Cohen, J. Mol. Biol. 274, 325-337 (1997).