Protein sequence analysis in silico: application of structure-based bioinformatics to genomic initiatives

David Michalovich*, John Overington† and Richard Fagan‡

The current pace of high-throughput genome sequencing programs coupled with high-throughput functional genomic screens has provided researchers with a bewildering array of sequence and biological data to contend with. Identification of proteins of interest from a particular biological study requires the application of bioinformatic tools to process and prioritise the data. From a protein function standpoint, transfer of annotation from known proteins to a novel target is currently the only practical way to convert vast quantities of raw sequence data into meaningful information. New bioinformatics tools now provide more sophisticated methods to transfer functional annotation, integrating sequence, family profile and structural search methodology. The importance of these approaches to medical research is increasing as we move to annotate the proteome through functional and structural genomic efforts.

Introduction

The advent of large-scale cDNA and genome sequencing projects has revolutionised how we identify novel genes and their encoded proteins. This, coupled with high-throughput experimental data from genetic, RNA expression and proteomic screens, has provided the researcher with an abundance of data to analyse and mine. The requirement for computational analysis of these high-throughput data sources is clear, and is now an established component of both industrial and academic genomic projects. One feature of large-scale experimental screens is that they often produce biologically interesting protein targets for which the biochemical function and/or structural fold assignment is unknown.

Historically, before the advent of large sequence databases, protein function would be determined through focused biochemical studies, the protein and cDNA sequences often being added as the last pieces of the functional jigsaw. We are now confronted with the opposite situation — where sequences are abundant, but functional information is scant. The main computational approaches use annotation transfer from a functionally defined protein sequence or structure to the novel query sequence. Accurate annotation transfer relies on two essential components: a well maintained and annotated source database, and computational methods to search and assign similarities from that database. This review focuses on how advanced protein annotation tools incorporating structure-based characterisation can be applied to genomic initiatives to enhance functional annotation and provide a driver for structural genomics initiatives.

Source databases of protein sequence and structure

Three major collaborative centres provide the main repositories for sequence data: the European Molecular Biology Laboratory (EMBL) data library located at the European Bioinformatics Institute, GenBank®, the National Institutes of Health database located at the National Centre for Biotechnology Information (NCBI) and the DNA Data Bank of Japan (DDBK; see Table 1 for URLs).

These three centres make up the International Nucleotide Sequence Database Collaboration, which feeds the protein databases such as SWISS-PROT™ and TrEMBL (translated EMBL) [1] and PIR (Protein Information Resource; [2]) (Table 1). SWISS-PROT™ is a highly curated database of protein sequence, derived from the EMBL nucleotide database. There is a minimal level of redundancy in SWISS-PROT™ data and each database entry is highly integrated with other bioinformatic reference sources. TrEMBL is a supplement of SWISS-PROT™ and represents the translations of EMBL nucleotide sequences not yet integrated into SWISS-PROT™. PIR-PSD (the PIR Protein Sequence Database) provides a similar resource to SWISS-PROT™ (Table 1).

The above databases have seen a large increase in the number of sequence entries over the past 10 years. Since its inception in 1986, the SWISS-PROT™ database has grown from around 5000 sequences to 107 000 entries. An even greater increase is seen in the nucleotide source databases such as GenBank, which contains 15 million entries (February 2002 statistics obtained from http://www.ncbi.nlm.nih.gov.GenBank.GenBankstats.html). However, the majority of sequences in the nucleotide databases are derived from high-throughput sources. As the number of unannotated sequences greatly outweighs the annotated...
Several annotated secondary databases can be viewed as sitting on top of the primary protein sequence sources. These contain computationally and expert reviewed multiple alignments and protein sequence patterns. They can be used to aid function annotation through the grouping of proteins into evolutionary families. PFAM (protein families database; [3]) and SMART (simple modular architecture research tool; [4]) are both multiple alignment based databases, whereas PRINTS (protein fingerprints database; [5]) uses sets of smaller motif alignments or ‘fingerprints’ to annotate novel proteins (see Table 1 for URLs). These secondary annotation systems have been combined into INTERPRO (integrated protein family to link annotation. Nominally, functional annotation becomes problematic, and more sophisticated tools are required. These use conserved features of a protein family to produce a statistical profile of the sequence and/or use structural features of the family to link annotation.

**Profile-based searching**

The most important single algorithmic development in sequence searching methodology has been the development of position-specific iterated BLAST (PSI-BLAST; [13]; see Table 1 for URL). This combines the speed of the BLAST pairwise alignment algorithm with the advantages of searching with a sequence profile. On the first iteration of the PSI-BLAST programme, a normal BLAST-type search is carried out. Sequences, which are found within a pre-determined significance threshold, are used to build a profile — an empirical description of the amino-acid residues found in homologues at each point of the query sequence. On subsequent iterations, the profile is used to search the database and the profile is refined based on the new matches identified. This method has the advantage that more distant sequence relationships can be found.

**Sequence-based search tools**

Protein annotation (excluding experimental determination of function) relies on comparison of the query sequence to a database of previously annotated sequences such as GenBank® and SWISSPROT. Commonly, this comparison is carried out using searching tools such as BLAST (basic local alignment search tool; [12,13]; see Table 1 for URL). Based on the matching sequences identified, annotation can be transferred to the novel sequences. This of course relies on the existence of a closely related biochemically annotated homologue. Nominally, functional annotation can be reliably transferred between proteins that share 30% sequence identity or above, confidence increasing with identity. Below 30% sequence identity, transferring functional annotation becomes problematic, and more sophisticated tools are required. These use conserved features of a protein family to produce a statistical profile of the sequence and/or use structural features of the family to link annotation.
However, care should be taken in the use of PSI-BLAST. One major consideration is the masking of the query sequence before a PSI-BLAST search. Because of the iterative process of PSI-BLAST, regions of low amino-acid sequence complexity can result in false matches being brought into the BLAST search. These can then corrupt future iterations of the search, bringing in false sequence relationships. This can be especially problematic for membrane-bound proteins, which contain transmembrane spanning regions of low sequence complexity (there are only so many hydrophobic amino acids to make up transmembrane domains and so, when pushed by PSI-BLAST, multi-transmembrane proteins may all start to match each other).

The NCBI's Conserved Domain Database search (Table 1) uses a variation of the PSI-BLAST algorithm called reverse-position-specific BLAST (RPS-BLAST; [14]). In this case, PFAM and SMART multiple sequence alignments are used to build a PSI-BLAST profile known as a position-specific score matrix (PSSM). The pre-calculated PSSMs are stored in the database and the query sequence is searched against this database [14].

Another advance is in the use hidden Markov models (HMMs) to produce a profile of protein family alignments. One of the attractions in using HMM representations of protein families is that they include additional information beyond amino-acid preference; in particular, position-specific gap penalties. There are several HMM based programs in use, such as HMMER [15], which is used in the generation of the PFAM and SMART databases [3,4], as well as the sequence alignment and modeling (SAM) family of algorithms [16].

Profile-based search tools, regardless of their specific implementation, require accurate alignments whether these are generated automatically or by hand. Because the 3-D structure of a protein is more strongly conserved than the primary sequence, using the 3-D structure information in building multiple alignments leads to the building of better profiles, and hence provides the best discriminators of homology [17,18]. This type of approach is embodied in algorithms such as 3D-PSSM and FUGUE (sequence-structure homology recognition resource), which can extend the identification of homologous relationships beyond the scope of a PSI-BLAST-based search [19,20] using structurally defined alignments.

**Structure-based annotation**

As discussed, 3-D structure information can be usefully integrated in the process of identifying distant homologues. Clearly, structural information can be used to improve alignments. However, structure can also be used as a template on which protein sequence can be matched and validated through a variety of techniques. This fitting or ‘threading’ of a protein sequence onto structure often uses empirically defined energy potentials derived from alignment of sequences and structures. The first algorithm to employ this technique was THREADER [21]. THREADER can try to predict a fold of a protein even when there is little or no sequence identity, using a large library of protein folds as its database. However, there is no assessment of significance of the match, requiring careful expert interpretation of the data. Practically, the programme is slow and compute-intensive.

To make the algorithm more accessible and reliable, it has undergone several revisions since its inception, which have improved its processing speed and provided a more robust statistical scoring scheme, making it more applicable by the non-specialist [22]. Now termed GenThreader, the algorithm has evolved to be significantly different from THREADER and encompasses aspects of profile-based searching in conjunction with structural alignments [22]. When Jones et al. applied it to annotation of the *Mycoplasma genitalium* genome, they were able to assign a 3-D structure to at least one domain of 46% of the open reading frames [23].

**Application of protein annotation to a functional genomics target discovery project**

High-throughput functional genomics methods such as transcript profiling, protein interaction mapping and large-scale phenotypic analysis have been applied to many biological systems to decipher the full complement of genes or proteins that play a potential role in a pathway or molecular response. As a result, hypotheses of biological function are available for many previously uncharacterised genes where the biochemical or molecular function is still unknown. Although no single high-throughput method can unequivocally define gene function, combining the data obtained from any of these complementary approaches is likely to provide greater functional insight [24].

At Inpharmatica (London, UK) we have taken modified versions of GenThreader (GenomeThreader™) and PSI-BLAST and applied these to create an all-by-all protein comparison at an industrial scale. The sequence and structural relationships are stored and presented in a bioinformatic solution termed the Biopendium™. Biopendium™ allows us to investigate sequence and structural protein relationships not possible by single database queries.

Using high-throughput functional data in conjunction with the Biopendium™, we can put the highest current level of annotation, at both the biological and biochemical level, on pathways of interest. There are many examples in the literature where a protein interaction mapping experiment or transcriptome experiment has yielded interesting results, but the molecular or biochemical function of the selected proteins is not known. Knowing the molecular function of such proteins can allow one to reach a judgement as to the meaning of the data and, more importantly to the pharmaceutical and biotechnology sector, design assays against the target or further validate the target.
Chabas et al. [25] have used a functional genomics approach to study multiple sclerosis (MS). They carried out large-scale parallel sequencing of cDNA libraries that were derived from plaques dissected from brains of patients with MS and control brain samples. The results indicated a relative abundance of transcripts for osteopontin, a T-helper cell (Th) cytokine. The authors were able to conclude from this and several other pieces of data that osteopontin appears to regulate Th-1-mediated demyelinating disease, and that it may offer a potential target in blocking development of progressive MS. Osteopontin was not the only protein that met their filtering requirements; 53 proteins in total were found to be upregulated in the MS plaque cDNA libraries. Of the 53 proteins on their list, 15 carried little or no biochemical annotation. These 15 proteins represent 28% of the MS-specific proteins and include the proteins CGI-49, KIAA0517 and KIAA0582. We have used the Biopendium™ to put structural annotation on this dataset. Of the 15 biochemically sparsely annotated proteins, we could add structural and biochemical annotation to 11. Of the 53 proteins in total, we were able to add structural annotation to 70%. A representative example of this process is the CGI-49 protein.

The CGI-49 protein was identified by Lai et al. [26] as a human gene that is evolutionarily conserved in Caenorhabditis elegans. This is essentially all that is known about this protein in the public domain; specifically, the biochemical and molecular function of CGI-49 is unknown. Chabas et al. [25] now identify that it is upregulated in MS plaques and not in normal brain. Hence, like osteopontin, CGI-49 may offer itself as a potential target in blocking the development of MS. The Biopendium™ annotates CGI-49 as a sugar hydrolase (NADP-binding Rossman fold; Figure 1). The sequence identity between CGI-49 and the representative sugar hydrolases is approximately 16%. By using the Biopendium™, we can delve further into the biochemical function of CGI-49 (Figure 2). We see here that the residues involved in sugar binding as well as those involved in NADP binding in the sugar...
reductase (taken from the structure of saccharopine reductase, PDB code 1ESQ) are conserved in CGI-49. With this information, we conclude that CGI-49 has a fold similar to the sugar hydrolases and will bind NADP and a sugar, and hence function as a sugar hydrolase. With this information, a biologist could now develop an assay to test this and run a screen to look for small-molecule inhibitors of this protein.

It is interesting that in the list of the 53 proteins identified in the study by Chabas et al. [25], two proteins known to be involved in sugar metabolism were identified: lactate dehydrogenase and phosphoglycerate mutase. Using the Biopendium™, we are able to identify another protein involved in sugar metabolism as being present in these MS brain plaques. This suggests that sugar metabolic pathways may be deregulated or function anomalously in these MS plaques. Biologically, this may be relevant in that sugar homeostasis must be controlled within narrow limits in the brain. From this study using the Biopendium™, we were able to place proteins into pathways which were known to be relevant or playing a part in the MS plaque pathology.

Linking sequence space to structure
As discussed previously, large-scale structural genomic initiatives are already being embarked upon. Their goals range from determining representative structures for all protein folds, through to detailed structural analysis of the products of all 35 000 or more human genes [27,28]. Structural genomic programmes are therefore likely to have a huge impact on our understanding of protein function. When these initiatives are married with structural bioinformatics tools, we should be in a strong position to annotate whole proteomes through such approaches. First, however, it is important to estimate the current structural coverage of known proteins and to what depth we can structurally annotate known proteins using bioinformatic tools. It has been estimated that it would take around 16 000 selected structures to provide 3-D models for the
vast majority of proteins [29••]. In this study the authors applied bioinformatic approaches to estimate the accuracy of structural coverage of the known proteins and extrapolated their results to large-scale studies.

Given that approximately one in 10 structures currently deposited in the PDB database represents a novel fold (http://www.rcsb.org/pdb/folds_table.html), future selection of candidate proteins for structural studies should take into account whether the protein is likely to represent a novel fold, and therefore add to our structural coverage, or simply add another member to an existing fold class. To that end, maximising the ability to annotate existing folds is of great importance. Another problem to overcome is the fact that some folds are promiscuous in function — for instance, the βα TIM-barrel fold is found in several functionally distinct protein families [30••, 31] — therefore, fold assignment alone may not be enough for our goal of determining biochemical function. We need to aim for depth as well as breadth of coverage; having a range of tools that allows us to move between structural and sequence annotation should allow us to achieve this.

SCOP and other structural classification systems provide an excellent point for linking sequences and structure. Aloy et al. [30••] have used SCOP to link sequence alignments from PFAM and SMART to structures classified in SCOP. The authors were able to detect superfamilly relationships using structure-based sequence alignments that were not previously determined in the starting databases. This approach helps close the gap between superfamilies, providing an improved annotation system for placing new structures into protein folds as well as for annotating novel sequence.

Conclusions
How far can we infer biochemical function based on structural similarity? Clearly, confidence of annotation grows with homology. It is estimated that if we can decrease the accuracy threshold required to assign structure to sequence from 30% to 20% using bioinformatic tools, we would reduce the number of experimental structures required to cover the proteome by half [29••]. As with many bioinformatic approaches, it is the combination of tools and data resources which provides the best results. Integrating sequence-based, profile-based and structure-based search algorithms in systems such as Biopendium™ coupled with the increase in structural information available will greatly enhance our ability to annotate novel sequences. This, in turn will extend our ability to prioritise targets derived from functional genomic studies.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


