DDT Vol. 7, No. 9 May 2002

Prioritizing the proteome: identifying pharmaceutically relevant targets

Mark B. Swindells and John P. Overington

Considerable attention is now being placed on prioritizing the proteome as the point of delivery for genomic information. Some of the challenges faced in prioritizing efforts from a pharmaceutical perspective, when presented with an incomplete proteome picture, are described. Examples of pharmaceutically relevant proteins are used to illustrate an informatics-based analysis of the proteome using knowledge of known drug targets. We show how results can be maximized by linking informatics approaches to experimental techniques and describe methods that can be used for prioritization within unprecedented protein families using, for example, single nucleotide polymorphism data and knowledge of disease pathways.

Mark B. Swindells* and John P. Overington
Inpharmatica
60 Charlotte Street
London, UK W1T 2NU
*tel: +44 20 7074 4600
fax: +44 20 7074 4700
e-mail: m.swindells@inpharmatica.co.uk

The human genome [1,2], as well as those of many other eukaryotes [3–5], bacteria [6,7] archaea [8] and of course viruses [9] (sample references only, see genome pages at http://www.ncbi.nlm.nih.gov for comprehensive lists), have now been sequenced, providing us with unprecedented 'piece lists' for these organisms. With this volume of DNA now available, it is often assumed that related problems, such as assembling the DNA into continuous regions, will melt away and that more aspirational challenges, such as folding three-dimensional (3D) structures from sequence [10], are closer to being addressed.

From genome to proteome

The reality, however, is some way from this. Although the human genome was sequenced in double-quick time, the challenge of comprehensively identifying the constituents of the human proteome has only just begun. From the progress with simpler organisms this is also likely to be slow. Five years after the yeast *Saccharomyces cerevisiae* genome was completed (the first completely sequenced eukaryote), the identification of its proteome remains an ongoing task with many open reading frames (ORFs) remaining hypothetical or of unknown function [11], and the ongoing discovery of novel ORFs.

The reason for stating these challenges up front is to set the context for the developments that we (these authors included) often describe as great progress for the field. Prioritizing the human proteome for pharmaceutically relevant targets is going to be considerably restricted by the fact that, at present, nobody has a complete view of that proteome. Nevertheless, it is possible to make real progress by reapplying knowledge from previous successes to present-day challenges.

Protein families as drug targets

There are now many launched drugs and, with effort, it is possible to collate both the drugs and their cognate targets. The first comprehensive collation of drug targets was published by Drews [12], who estimated that there were ~500 targets for launched drugs (including protein therapeutics). In fact, the number could be significantly lower because some targets appear more than once, hiding behind different names. There is also the complication of identifying precisely which structural domain of a protein is involved in binding a drug. In HIV-reverse transcriptase (RT) there are distinct binding sites for the two mechanistic classes of RT inhibitors. For instance, AZT binds at the nucleoside site, whereas Efiravenz binds to the non-nucleoside allosteric site. All this information is important for subclassification and the correct application of these data to analogous systems.

Through hand collation of the targets for all currently approved US pharmaceuticals, starting from several key sources [13–15], we find that enzymes represent the largest class of targets for drug action and cover ~50% of targets for launched drugs. Many of the

1359-6446/02/$ – see front matter PII: S1359-6446(02)02250-X

enzymes, such as neuraminidase, HIV proteinase and thrombin, belong to distinct structural families (β-propellers, aspartic proteinases and trypsin-like-serine proteinases, respectively). By contrast, the second most common functional class, ligand receptors, is dominated by the G-protein-coupled receptor (GPCR) subfamily of seven transmembrane (7TM) proteins. As a result, although it is reasonable to concentrate on GPCR family homologues to identify additional target members because of their historical dominance on a per family basis, there are many other precedented families (some of which receive little attention) that could offer valuable returns on further identification and study. Most reported *in silico* and experimental target discovery efforts have typically focussed around these highly precedented families.

A successful drug needs to bind to a target that is a valid point of intervention for a disease, be ideally specific to that target and be safe (i.e. any other binding or biotransformation that does occur will not endanger the patient). This, in practice, is a lot to worry about when one does not even know the full complement of the proteome. Nevertheless, computational approaches, when combined with appropriate high-throughput experiments, can be surprisingly efficient at making inroads to this problem. The following text describes approaches that are now used to tackle these issues.

Expression arrays

A popular way to prioritize the proteome is to use expression arrays to identify transcripts of proteins that are differentially expressed in specific cell types [16–18]. Based on an assumption that some of these could have a causative role in the disease under investigation, bioinformatics subsequently has an important role in further prioritizing the differentially expressed proteins, as there will still be too many to fully validate by experimental means. Therefore, by concentrating on proteins that fall into families whose members bind to known drug molecules, it is possible to select a subset for initial analysis. If these proteins subsequently fail to yield interesting results, one can further prioritize the remaining transcripts using less rigorous requirements.

From an informatics view, the first problem is defining domain families with drugged members. Currently, this can only be achieved manually. On a small scale, a pragmatic approach is to reconcile current in-house abilities and historic group or company successes with the list of targets originally collated by Drews.

The second problem, placing proteins into these selected families, can be assisted by improvements in technology. Until relatively recently this was mainly done with Blast searches [19] and other techniques based on pairwise alignment. However, many additional homologues fall beyond the reach of a pairwise search because these simple search methods are unable to detect homology when the pairwise sequence identity falls below ~25–30%. Fortunately, this can be easily improved by applying profile-based searches, such as those possible with PSI-Blast [19] or, alternatively, pre-generated profiles known as Hidden Markov Models [20], which are used by Pfam [21]. For more expert use it is possible to push the limits even further for those targets with 3D structure information by apply threading based methods [22,23]. With such threading methods it is frequently possible to identify relationships with less than 15% amino acid identity and, because there are potentially many more distantly related proteins than closely related ones, there are practical benefits to applying such technologies [24]. If the human proteome were completely mapped, these methods would work extremely well because far more protein domains can be mapped onto well-characterized proteins than were previously envisaged.

However, this is not the case and our current view of the proteome consists of:

- A relatively small (a few thousand) number of well-characterized proteins.
- A larger number of cDNA sequences for which the protein coding region can be reliably inferred.
- A large set of homology 'confirmed' statistical predictions (such as a Genscan [25] ORF prediction with a region of homology to a well-characterized protein).

This final set can also be subdivided on the basis of other supporting information, such as expressed sequence tag (EST) hits and the presence of upstream regulatory elements and intron and/or exon conservation. But even this will be incomplete as it is well known that statistical prediction techniques miss well-characterized proteins and, therefore, by extrapolation, will fail to pick up novel real proteins. For instance, in the current release of the Golden Path (website at the University of Santa Cruz, CA, USA; http://genome.ucsc.edu), rather than predicting the FIL-1 zeta gene (an interleukin-1 homologue), Genscan predicts an ORF on the opposite strand.

The only ways to address this problem are to either make larger numbers of predictions (e.g. by combining different algorithms) or to avoid any statistical prediction of ORFs and simply search for homologies directly in the DNA. However, there are problems with both of these approaches. The former, although used to good effect at many centres (e.g. see http://genome.ucsc.edu) is still unlikely to provide a comprehensive set of genes. The latter has several issues that make it more difficult to implement successfully:

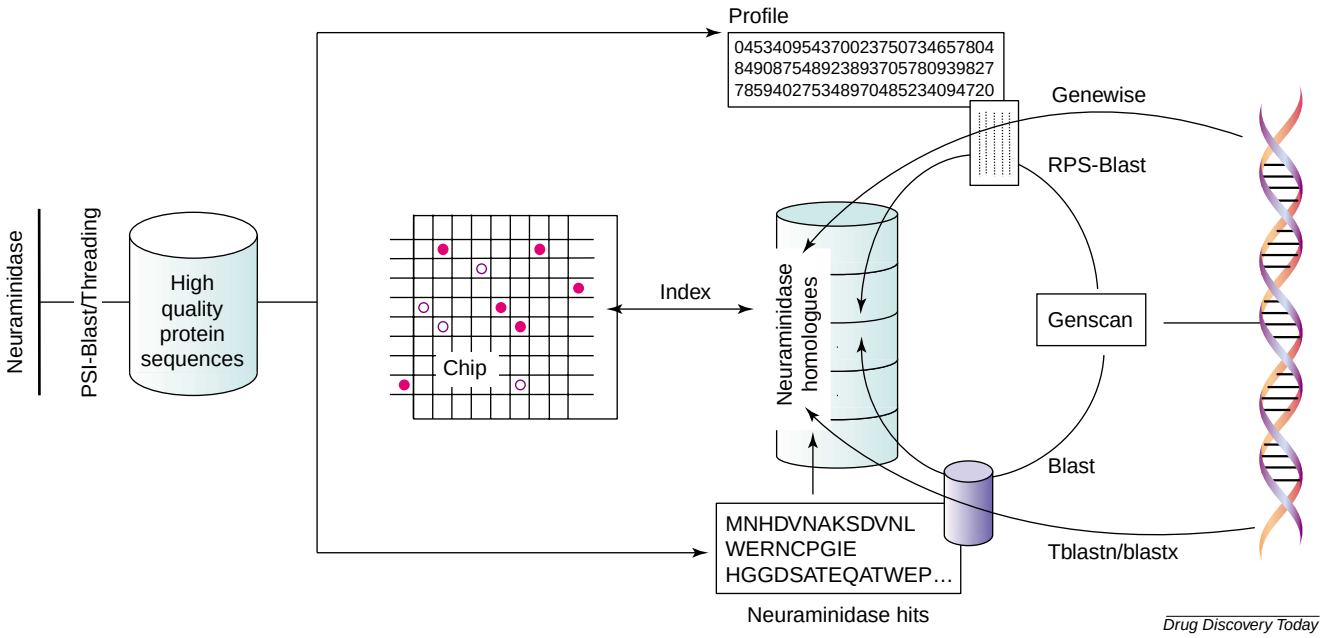

Figure 1. Structure-based drug design. Using neuraminidase as an example, a set of close and distant neuraminidase homologues (hits) and profiles can be identified by comparing the neuraminidase to a protein sequence database using PSI-Blast and threading algorithms. Genewise and Blast family programs can then identify additional homologues from the genome through direct DNA and/or protein comparisons. Finally, with this comprehensive list, standard chips can be annotated for the presence of (in this case neuraminidase) homologues or custom chips can be designed.

- It requires extensive computational time because ~100-times more data needs to be searched.
- Human proteins are typically divided into a series of exons, the 'signal' that guides the homology search to relevant hits is considerably weakened.
- The DNA of a chromosome is too large for standard computers and algorithms to manage in one chunk, it must be divided into sections. Each analyzed section must therefore overlap with its neighbours, so that exons from the same gene have the chance of appearing together.

There are now several programs with which to conduct these searches. The simplest is, again, a Blast program, in this case blastx, which compares a protein sequence against a database of DNA sequences.

Because of the limitations described previously, these methods will probably be unable to reliably identify relationships below ~70% amino acid identity. For more sensitive searching, other Blast tools can be used to compare DNA against a profile generated by PSI-Blast. For a potentially more accurate and sensitive search, Genewise [26,27] can be applied to a Hidden Markov Model, which has the advantage of encoding additional information about the gene model. However, these searches are exceptionally CPU (central processing unit) intensive. Even refining a gene prediction over a 2MB section of DNA can take ~3 hours on a standard Linux workstation. On this basis, to search just one protein against the complete genome (~3GB) would take over half a year on a single machine. One would either need to be patient or have enough money to purchase a large Linux farm or dedicated hardware, such as those developed by Paracel (Pasadena, CA, USA; http://www.paracel.com).

One of the main practical challenges in mining these data is assembling the algorithms into a process that can generate useful results. Such pipelines tend to be complex but Fig. 1 summarizes how relatives of a specific target protein or domain could be identified using the algorithms introduced in this article.

Over the coming year we anticipate progress to be made through comparative genomics, specifically mapping the mouse genome onto the human. At least in the public domain, the mouse sequence has not been completed to the accuracy of the human. At present, ~10% is considered suitable for extensive use (http://mouse.ensembl.org). At sites such as the Golden Path (http://genome.ucsc.edu) and Ensembl (http://www.ensembl.org) the fruits of these initial mappings can be viewed together with a variety of other data covering many of the topics already discussed here. The advantage of the combined human–mouse mapping is that, in line with many earlier observations, genes will be broadly conserved between the two organisms and the exonic regions will be much less variable than the

corresponding introns. In this manner, the noise that previously came from not being able to place exons accurately is considerably reduced. In addition, it could even be possible to start tying adjacent exon predictions together to strengthen the signal and facilitate detection by previously described techniques.

Linking experiments with informatics

It could be that the best results for prioritizing the proteome will come from a strong interplay between experiment and informatics. In Fig. 1 we show how the various informatics approaches described so far fit with experiments using neuraminidase as a well known example of structure-based drug design [28]. However, it would be relatively easy using the same approach to extend the application into different drug families or, alternatively, to further increase coverage over the neuraminidase superfamily (distantly related proteins) and fold (proteins that have the same 3D shape as neuraminidase but might not be related or have a similar molecular function). This second approach is particularly dependent on knowledge of the 3D structure. However, a good result can be achieved relatively easily by taking advantage of our knowledge of relationships identified through 3D structure alone. For instance, the cysteine knot family of growth factors have surprisingly similar 3D structures even though their sequences are considerably different [29]. By selecting appropriate structures, for example nerve growth factor (NGF), transforming growth factor-β2 (TGF-β2), gonadotrophin and plasminogen-derived growth factor (PDGF), a far more impressive set of sequences will be annotated than through a single protein. Even for proteins that are more amenable to profile-based searching, such as the protein kinases, better coverage will be achieved by selecting a basket of starting proteins. A particularly good example is the atypical protein kinase domain of a TRP channel whose 3D structure was recently determined by Kuriyan's group [30]. All but the most sensitive algorithms fail to identify this as being structurally similar to the more familiar protein kinases on the basis of sequence alone.

Disease and pathways

The approaches that we have described so far combine experimental prioritization of probable disease targets with *in silico* predictions of the pharmaceutically relevant proteins based on precedent. However, there are other ways in which a knowledge of bioinformatics and, in particular, 3D structure can be applied to prioritization. Two contemporary examples are; (1) identifying single nucleotide polymorphism (SNP) data related to disease (rather than background polymorphisms) and (2) annotating MS data.

SNPs

SNP data are now being collected on a large scale both in the pubic domain and privately. When a general association with disease is being sought it is difficult to separate variations of key importance with non-crucial variations within the population. Two methods are currently popular. The first seeks to identify polymorphisms in the control regions of proteins rather than the proteins themselves on the assumption that these could have an important effect on expression [31]. These approaches require a good view of gene structure, which – as we have already described – is itself far from a trivial challenge. The alternative is to prioritize those in the protein-coding regions and for this 3D structure has a key role. If a protein contains a coding SNP that is not silent (i.e. the amino acid also changes) it is possible to predict its probable impact if the 3D structure of a homologue (however distant) is available. For instance, if the SNP alters an amino acid at a position that is either buried, near to a known active site or ligand binding site, or contributing to a protein–protein interaction region, then it is more likely to modify that protein's activity than if it were at another position. There are two particularly clear examples of this in the literature where amino acid mutations lead to hypertension [32] and diabetes [33].

MS data

The technique of MS is increasingly being targeted to the elucidation of specific pathways. A particular example related to our own work is the Alzheimer's disease-associated presenilin–β-amyloid precursor protein complex, where MS identified a protein now known as Nicastrin [34]. Our subsequent application of bioinformatics, particularly 3D structure-based threading techniques, indicated that Nicastrin was a member of the aminopeptidase–transferrin receptor superfamily [35]. Most recently, the same group has also identified Nicastrin binding to membrane tethered notch [36]. Here the purpose again is to annotate as many of the proteins identified by MS and, ideally, to find pharmaceutically relevant components. However, when investigating such pathways, it is important to annotate all of the proteins at a molecular level and when the number of interaction partners is large this becomes a time-consuming task. As a result, the search methods introduced earlier must be efficiently scaled to cover as many protein families as possible.

Scale-up

The extent to which all these approaches are scaled depends on the value attached to comprehensive coverage. For instance, the approaches described in Fig. 1 could essentially be applied by any knowledgeable bioinformatician with access to the appropriate software. Scaling-up to

cover all proteins of known 3D structure, however, would be more of a challenge because the amount of data produced would require systematic calculation and storage, the use of technologies such as relational databases, the application of a small farm of Linux based machines to calculate and recalculate as new data become available, as well as the associated problems of queuing and monitoring jobs on such a system. By now, for anyone doing the work seriously, this has already become a full-time job for about five people. However, there are many proteins that will not fit into these families. For these to be annotated, essentially all proteins would need to be sent through parts of these search protocols and additional data resources would need to be added. This is essentially the approach that we have adopted at Inpharmatica (London, UK), where each unique protein sequence is applied equally to the relevant algorithms available. To achieve this, ~20 people are now required and one of the largest Linux clusters in the world – and even this number excludes work on any enhancements that might be made to move the system forward.

For Nicastrin we were essentially able to achieve our insight [35] in 10 minutes because of the previous investment in high-throughput annotation technologies. However, more important than the time spent, only the more sophisticated approaches described would have been able to achieve the result. The primary reason is that the transferrin receptor superfamily is not typically considered to be a pharmaceutically relevant family. Whether Nicastrin subsequently turns out to be a good target for pharmaceutical research remains an open question. However, it is true that Nicastrin, at the time of publication, was the rare case of a new mechanistically relevant target in a therapeutic area crowded by patents. Although it could be argued retrospectively that the same annotation might have been found by an expert, the facts remain that:

- There is no published annotation that precedes our own.
- With the high volumes of experimental data now being generated, only a vanishingly small subset will be analyzed by experts working in the traditional vein.

Moving forward

Therefore, the challenge moving forward will be to make as much use of data as possible. There are now many experiments that have been successfully scaled-up from the lab experiment to industrial scale. The most famous will always be DNA sequencing but now MS and even 3D structure-determination is being addressed in ways that only a few years ago would have been thought unrealistic. To link these data resources in an intelligent manner requires much more than the often-touted integration. It requires the ability to build significant knowledge from these individual databases so that high quality observations can be found more often than previously. This will involve expert bioinformaticians of the traditional mould, experimentalists and computer engineers working together and taking advantage of each other's skill sets. If the appropriate efforts are taken, it will be possible in a relatively short time to prioritize protein targets comprehensively using data from a variety of sources, thereby consigning articles such as these to the history books.

Acknowledgement

Marlon Schwarz played a key role in manually curating drug targets.

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
2 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921
3 The *Caenorhabditis elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
4 Adams, M.D. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195
5 *Nature* (1997) The yeast genome directory. *Nature* 387 (Suppl. 5)
6 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science* 269, 496–512
7 Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544
8 Bult, C.J. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073
9 Sanger, F. *et al.* (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695
10 Dahiyat, B.I. and Mayo, S.L. (1997) *De novo* protein design: fully automated sequence selection. *Science* 278, 82–87
11 Gavin, A.C. *et al.* (2002) Functional organisation of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
12 Drews, J. (1986) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14, 1516–1518
13 Goodman and Gillman's *The Pharmacological Basis of Therapeutics*. (1995) (9th edn), (J.G. Hardman, ed.), McGraw-Hill
14 Sneader, W. (1996) *Drug Prototypes and their Exploitation*. Wiley
15 *Therapeutic drugs.* (1999) (C. Dollery, ed.), Churchill Livingstone
16 Zweiger, G. (1999) Knowledge discovery in gene-expression microarray data: mining the information output of the genome. *Trends Biotechnol.* 17, 429–436
17 Strachan,T. *et al.* (1997) A new dimension for the human genome project: towards comprehensive expression maps. *Nat. Genet.* 16, 126–132
18 Khan, J. *et al.* (1999) DNA microarray technology: the anticipated impact on the study of human disease. *Biochim. Biophys. Acta* 1423, M17–M28
19 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
20 Karplus, K. *et al.* (1999) Predicting protein structure using only sequence information. *Proteins* 37, 121–125
21 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266
22 Jones, D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature* 358, 86–89

23 Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815

24 Todd, A.E. *et al.* (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* 307, 1113–1143

25 Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346–354

26 Guigo, R. *et al.* (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* 10, 1631–1642

27 Birney, E. and Durbin, R. (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* 10, 547–548

28 Taylor, N.R. (1998) Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J. Med. Chem.* 41, 798–807

29 Swindells, M.B. (1992) Structural similarity between transforming growth factor-β2 and nerve growth factor. *Science* 258, 1160–1161

30 Yamaguchi, H. *et al.* (2001) Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity. *Mol. Cell.* 7, 1047–1057

31 Suthanthiran, M. (2000) The importance of genetic polymorphisms in renal transplantation. *Curr. Opin. Urol.* 10, 71–75

32 Geller, D.S. *et al.* (2000) Activating mineralocorticoid receptor mutation in hypertension exacerbated by pregnancy. *Science* 289, 119–123

33 Barroso, I. *et al.* (1999) Dominant negative mutations in human PPAR gamma associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* 402, 880–883

34 Yu, G. *et al.* (2000) Nicastrin modulates presenilin-mediated notch/glp-1 signal transduction and βAPP processing. *Nature* 407, 48–54

35 Fagan, R. *et al.* (2001) Nicastrin, a presenilin-interacting protein, contains an aminopeptidase/transferrin receptor superfamily domain. *Trends Biochem. Sci.* 26, 213–214

36 Chen, F. *et al.* (2001) Nicastrin binds to membrane-tethered Notch. *Nat. Cell Biol.* 3, 751–754