Recent developments in computational proteomics

Edward T. Maggio and Kal Ramnarayan

The mapping of the human genome was completed earlier this year and efforts are underway to understand the role of gene products (i.e. proteins) in biological pathways and human disease and to exploit their functional roles to derive protein therapeutics and protein-based drugs. A key component to the next revolution in the 'post-genomic' era will be the increasingly widespread use of protein structure in rational experimental design. Improvements in quality, availability and utility of large-scale three- and four-dimensional protein structural information are enabling a revolution in rational design, having particular impact on drug discovery and optimization. New computational methodologies now yield modeled structures that are, in many cases, quantitatively comparable with crystal structures, at a fraction of the cost.

Edward T. Maggio*
Kal Ramnarayan
Structural Bioinformatics,
San Diego, CA 92127, USA
*tel: +1 858 675 2400
fax@ +1 858 618 1040
e-mail: info@strubix.com

▼ With the understanding of chemical structure in the early 1800s, the elucidation of atomic structure in the early 1900s, the comprehension of polymer structure in the 1920s, the determination of the first protein structure in 1957 and, more recently, the determination of gene structure in the 1970s, structural knowledge has led the way to every major commercial revolution in the chemical and physical sciences over the past two centuries. At the beginning of this new millennium, computational proteomics – the large-scale generation and analysis of three- and four- dimensional (3D and 4D) protein structural information and the application of structural knowledge across all life science disciplines – promises a scientific and commercial revolution of unequaled impact on mankind.

Structure enables!

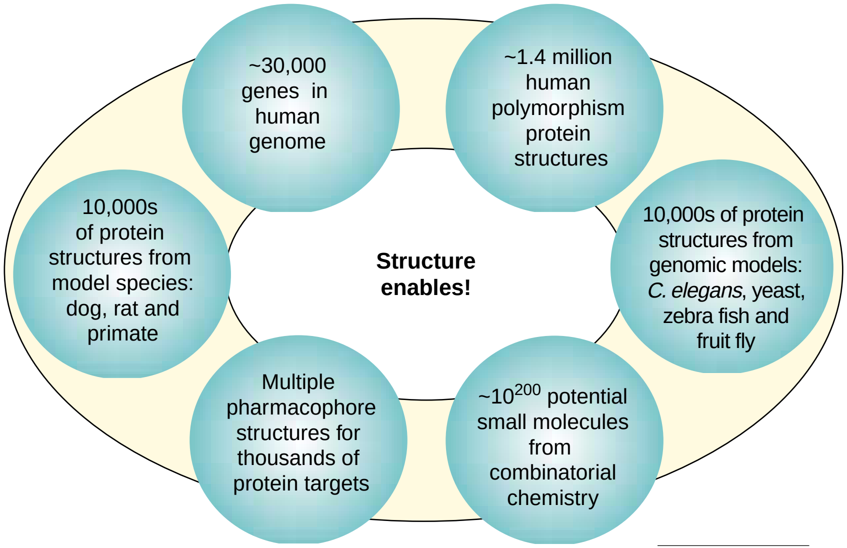
In a single snapshot, Fig. 1 presents not only the problem but also the opportunity that life scientists now face – particularly scientists in the pharmaceutical industry. The three billion base pairs in the human genome have been sequenced and provide a 'working draft' of the entire genome[1,2]. Although all the drugs available today work on only ~500 unique targets, the human genome project has revealed the sequences of ~30,000 unique genes – any one of which is a potential new drug target. By some estimates there could be 10 to 100 slight variants (genetic polymorphisms) for each target gene. Among the pharmacological model species, such as the dog and rat, and the genomic model species, such as yeast and *Caenorhabditis elegans*, there are certainly tens of thousands of sequences – and therefore structures – of potential interest to life scientists. Combining this with the estimated 10^{200} potential small molecules accessible through combinatorial chemistry makes one opportunity clear; numerous drug targets and their corresponding drugs will emerge from the human genome project and other genomic studies.

Although most of the genome is now available as sequence data, little is known about protein function and, in turn, even less is known about individual protein structures. Because access to high-quality protein structural information fundamentally enables rational experimental design in essentially all life-science disciplines – from molecular biology, molecular pharmacology, medicinal chemistry and structure-based drug design, to the many sub-disciplines of computational biology – the major challenge of the post-genomic era will be to translate all of this sequence data into structural knowledge. Structures will be the key to unlocking the coming post-genomic revolution in human medicine, in understanding, diagnosing, preventing and treating human diseases. This fact is implicitly acknowledged and confirmed by the scientific community's commitment to a growing number of large-scale 'structural proteomics' or 'structural genomics' projects

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(01)02003-7

in the USA (Refs 3,4), Europe[5] and Asia[6]. For example, the National Institutes of Health recently initiated a US$125 million 'Protein Structure Initiative' that started in late September 2000 and is aimed at producing 10,000 X-ray crystal structures, to be made available on the Internet. Also, the initiative entitled 'Protein Folds Project' is being established at the Institute of Physical and Chemical Research (RIKEN, Yokohama, Japan), in which an assembly of 16 state-of-the-art high-resolution NMR spectrometers will be used to determine the 3D structures of smaller protein molecules on a scale previously not imagined.

Such efforts, whether government-funded or arising from the recent spate of private companies springing up in direct competition with these government efforts, rely on attempts to scale-up and accelerate physical methods of structure determination, such as X-ray crystallography. However, the principal hurdles of cloning, expression, purification and crystallization remain largely idiosyncratic processes that are not easily subject to brute-force acceleration, and account for the fact that only one in 20 proteins ultimately yields useful crystals for structure studies. Some proteins, such as membrane-associated proteins, might not crystallize at all. Although it might appear to be a rather questionable endeavor for private companies to compete with the growing number of well-funded government structure determination initiatives, all such activities, irrespective of the source, are scientifically valuable. The problem posed by the enormous number of targets is not subject to a 'bigger hammer' approach; merely applying more resources to the increased number of targets without significantly improving efficiency is not only inelegant but impractical. Today, a computational proteomics approach based on protein structural modeling is the only practical approach in sight.


~30,000 genes in human genome
~1.4 million human polymorphism protein structures
10,000s of protein structures from model species: dog, rat and primate
Structure enables!
10,000s of protein structures from genomic models: *C. elegans*, yeast, zebra fish and fruit fly
Multiple pharmacophore structures for thousands of protein targets
~10^{200} potential small molecules from combinatorial chemistry

Drug Discovery Today

Figure 1. A schematic representation of information and data flowing into the discovery environment. Now that the mapping of the human genome has been completed, the number of drug targets has increased tremendously. Analysis is further complicated by the number of polymorphisms that can potentially arise. In addition, there is a wealth of information from sequencing projects of model species and organisms in which there is enormous opportunity for protein structure analyses. The discovery landscape is further overloaded with information by multiple pharmacophore definitions for the new targets and the possibility of accessing enormous numbers of chemicals through combinatorial chemistry.

Will modeled protein structures replace crystal structures?

The field of computational proteomics stands squarely on the shoulders of X-ray crystallography. Homology modeling is limited by the need for at least one crystal structure within each fold-class to be modeled. Experimentally determined X-ray crystallographic protein structures within each particular protein-fold class make it possible to model up to hundreds of related proteins within the same class (determined using sequence or structural similarity). Fortunately, this limitation is being addressed aggressively. At present, there are more than 10,000 crystal structures in the publicly accessible Protein Data Bank (PDB, http://www.rcsb.org/pdb; Ref. 7). These structures, and the additional thousands of crystal structures that will be made generally available by the Protein Structure Initiative and other public or private efforts in the USA, Europe and Japan, will remain essential to the modeling process for the foreseeable future. X-ray crystallography is here to stay – at least for now! Through these efforts, we believe that the majority of all protein structures will be accessible in a high-resolution form using augmented homology modeling over the course of the next five years.

However, the aforementioned technical difficulties that relate to X-ray crystallography, along with the long time-frames and the significant costs associated with the determination of individual structures, make it extremely unlikely that X-ray crystallography will be the primary source of large-scale 3D protein structural information. Rather, technical improvements in X-ray crystallography and the relentless trends towards improved computational methodologies and ever-expanding computational horsepower will undoubtedly continue. Together, these elements of change will work synergistically to meet the challenges and promises of computational proteomics – large-scale availability, analysis and use of protein structural information for all life-science disciplines.

Although protein modeling is, quite literally, orders of magnitude faster and less costly than X-ray crystallography, two questions relating to quality and utility necessarily arise: (1) how 'good' are such modeled structures and (2) for what purposes are they useful? The answers to these fundamental questions require that we briefly recount the history and limitations of earlier modeling technologies, cite recent innovations that have gone a long way towards extending the boundaries of earlier modeling methods and provide examples of the current utility of modeled protein structures.

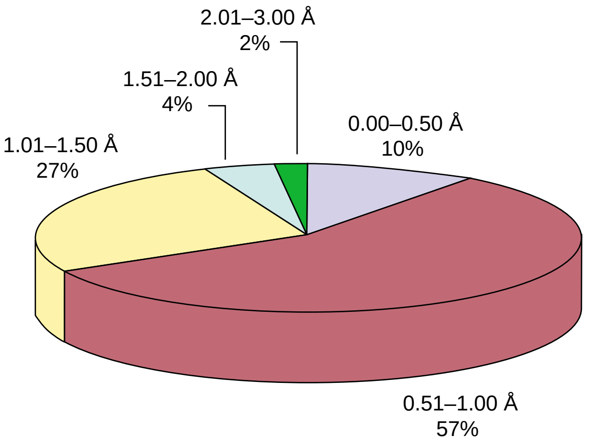
History of computational methods and current limitations

Homology modeling and threading technologies are often termed 'knowledge-based' methods. This means that scientists rely on structural knowledge of proteins for which 3D structures have already been determined to infer the structure of proteins for which only the sequence is known. Homology modeling to some extent, and threading to a much greater extent, are regarded as 'high-throughput and low-resolution' techniques. Historically, homology modeling has been regarded as yielding reasonably accurate core structures but relatively inaccurate surface structures. Unfortunately, most or all of the biological activity of proteins resides on the surfaces of the protein. In recent years, new computational methodologies have been developed to augment homology modeling to the point that augmented homology modeling can, in many cases, now yield structures quantitatively comparable with crystal structures.

Threading methodologies are employed when insufficient structural information is available to permit homology modeling. Rather than yielding a single predicted structure, threading methodologies yield multiple solutions to the protein prediction problem. It is not possible to determine if any single prediction represents the 'true' structure but it is assumed that the aggregate information in total reveals potentially useful perspectives on what the true structure of the protein might be. Even though predictions using threading approaches have progressed in the past few years, these approaches result in a predicted structure differing from the structure determined using X-ray crystallography by anywhere between 3.2 Å and 17.6 Å with respect to the root mean square (rms) deviation of the protein C^{α} trace[8]. To put this in perspective, two crystal structures of the same protein would typically differ from 1.5 to 1.8 Å. Threaded structures are, thus, well outside the useful range for structure-based drug design but such structures are finding use in efforts to predict protein function at a general level.

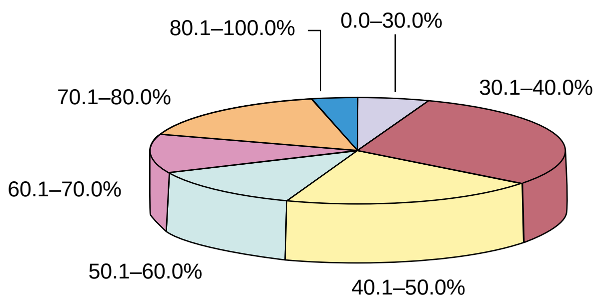
NMR determination of protein structure using Nuclear Overhauser Effect (NOE) intensities (distance constraints) combines physical distance determinations with computationally intense refinement techniques and, thus, might arguably be included in a discussion of computational methodologies. As we move up to increasingly larger target molecule size, the processing of NMR measurements becomes computationally more burdensome and is ultimately the limiting step in structure determination; however, exciting and significant computational advances are being made. Recently, Padilla and Karlov[9] described a rapid algorithm aimed towards processing large molecules with 1000 or more NOE intensities with an increase in speed of up to two to three orders of magnitude compared with state-of-the-art approaches. This development (termed ASTER for Automatic STructure Estimation and Refinement) promises to amplify the sheer output and absolute quality of protein structures determined using high-field NMR initiatives, such as the 'Protein Folds Project' at Japan's RIKEN facility.

Finally, as ever-faster computers melt away the barrier of computational speed, scientists will increasingly recognize, and come to grips with, the dynamic nature of protein structure. The 3D structures of proteins change rapidly over time, so that true protein structures are best represented as dynamic trajectories – a series of structures varying slightly with time, just as the images on a strip of motion picture film vary slightly with each frame. Such dynamic trajectories are, in effect, 4D protein structures. Because structure and function are intimately related to protein dynamic flexibility, 4D structure represents a new and untapped frontier of discovery and applications. Recently, the calculation of 4D structures has undergone a major advance with the development of a technology called Multi-Body Order(N) Dynamics [MBO(N)D], that enables computational scientists to simulate low-frequency molecular motions and properties, such as inter-domain movements, at least an order of magnitude faster than using conventional methods[10]. The mathematics underlying this advance was developed to model the dynamic motions of the various structural elements of large spacecraft. It has been shown that the essential dynamics are fully represented by sub-structuring of protein chemical groupings, such as residues or helices (individually called bodies), which aggregate the behaviors and properties of the atoms contained within each body, whether rigid or flexible. This makes the examination of motions over previously inaccessible long timeframes (i.e. >0.5–1.0 nanosecond) practical and opens the door to the routine use of 4D protein structure.



Drug Discovery Today

Figure 2. A pie chart showing the root mean squared (rms) comparisons between Structural Bioinformatics models and the subsequently determined crystal structures. For the 45 models used in the comparison, 96% were found to be within 1.5 Å or better and 98% were found to be within 2 Å.



Drug Discovery Today

Figure 3. A pie chart showing the break up of the percentage of homology of Structural Bioinformatics models used for the comparison with the crystal structures. Of the 45 structures, 35% of them have a sequence homology of 40% or less with their corresponding templates.

Continuing improvements in the quality of protein structural modeling

Over the course of the past decade, the process of homology modeling has improved significantly. Augmented homology modeling approaches have been developed that directly address the poor quality of predicted protein surface structural information. For example, several methodologies permit the accurate prediction of surface loops on protein structures on a true *ab initio* basis[11–15]. By computationally grafting highly accurate loop structures onto reasonable homology-generated core structures, making certain corrections in the core structure and conducting extensive refinement procedures, structures can often be generated that are comparable in quality with those obtained directly using X-ray crystallography. The use of increasingly accurate secondary structure predictions enhances sequence alignments, which are fundamentally important in generating high-quality structures.

The accuracy of models using the most advanced augmented homology methods now approaches X-ray crystallography resolutions (1.5–2.0 Å rms) for proteins having sequence identities as low as 30% with an appropriate fold-class template[16]. Figures 2 and 3 compare differences in the rms deviations between the backbone structures of 45 computationally derived structures found in Structural Bioinformatics' ProMax™ drug target database and the corresponding recently published crystal structures determined by scientists around the world and subsequently deposited and made available in the PDB.

As stated previously, two very high-quality X-ray crystallographic structures will typically differ in the rms deviation of their backbones by ~1.5 Å. For the 45 paired modeled and crystal structures, 96% were found to be within 1.5 Å or less and 98% were found to be within 2 Å. This compares with homology models generated by automated programs in which the rms deviations might be in the order of 4.5–6 Å or more, which is far outside the useful range in providing sufficient detail needed for drug discovery.

As the sequence homology between the modeled protein and the template structure of the appropriate fold class derived from X-ray crystallography is lowered, the accuracy of the model degrades. There is an effective low-end cut-off point for effective homology modeling of ~20% sequence identity. But even at 20% sequence identity the overall fold of a protein, which describes the arrangement of secondary structural elements comprising the protein in 3D, can often be determined. Such models are characterized by resolutions of ~5–10 Å rms – just overlapping the high-end of threading quality in the examples reported by Panchenko and colleagues[8]. Although this limits their use in structure-based drug design and library screening, such structures are comparable with, and often superior to, threaded structures. This is often sufficient to yield valuable information in predicting function of newly discovered proteins. And even at this lower end of sequence identity, some portions of the protein's structure might be highly conserved, such as the 3D arrangement of residues responsible for enzyme action. Such localized arrangements can be predicted with much higher accuracy and might sometimes be useful in generating hypotheses for computational library screening.

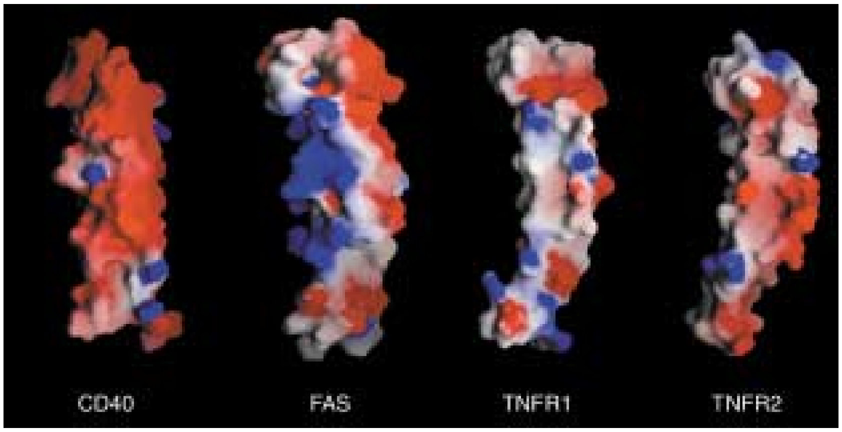


Figure 4. Electrostatic surface renderings of TNFR1, TNFR2, CD40 and FAS. Regions colored in blue indicate surface residues with a positive charge and those colored red indicate surface residues with a negative charge. Even though the receptors are structurally homologous, there are clear distinctions in the surface charge distributions for this important class of receptors.

Technical advances that will augment the performance of homology modeling and threading will continue to be made, steadily improving the accuracy of modeled structures. This is evidenced by the progress seen in methods to derive the secondary structure of proteins starting from their amino acid sequence[17–24]. Advanced neural network concepts developed at SBI-Advanced Technologies (Hørsholm, Denmark) now permit prediction of secondary structure from gene sequence with a previously unattainable accuracy of >80% making possible improvements in both homology modeling as well as threading[25]. Bohr *et al.* describe the development of structure prediction techniques incorporating advanced ‘Wring Mode’ concepts having further application in augmentation of homology methodologies[26]. Advances in both the generation and use of protein structures generated by threading are being made in several laboratories[8,27–29]. Mathematical techniques addressing the local minimum problem (in which the global lowest-energy structure is not obtained but rather a local low-energy one is found on minimization) in protein structure prediction promise to increase the usefulness of structures generated using both homology modeling and threading[9].

Increased availability and expanding utility of 3D protein structures

Computational proteomics in drug discovery

Comprehensive knowledge of protein structure affords a powerful advantage to modern drug discovery. Some of this advantage is afforded intrinsically by the structure and can be realized by simple examination. The identity of residues important to function or binding is tentative at best from sequence data alone because non-contiguous segments of the sequence are brought together in 3D during protein folding to form spatially contiguous sites. A simple examination of the structure resolves these difficult-to-impossible aspects of sequence analysis. If structures are available for orthologs (proteins that perform the same function in different species) or paralogs (proteins that have a common ancestry but that have evolved unique functions), differences in protein structure that lead to differences in function, specificity and modes of interaction can be understood. This knowledge can be crucial to an effective selection of pharmaceutical targets and is extremely difficult to develop using other methods.

The use of protein structural information in drug discovery underwent a significant change in the latter half of the 1990s. In the late 1980s, and through most of the 1990s, considerable effort was put into rational drug-design methods intended to design novel molecules that bind specifically to enzyme active sites or receptor-binding sites as elucidated in high-resolution crystal structures. Although these methods claim notable successes in optimizing drug leads, they have had a much smaller role in the initial discovery of lead compounds. The introduction of combinatorial chemistry and of HTS methods to the pharmaceutical industry has established a new discovery paradigm that offers significant advantages over X-ray crystallographic approaches in efficiency and cost. In the past few years, massively parallel methods have been developed that, often successfully, find potentially optimizable leads for new drug targets. However, the next generation paradigm, which is an entirely new structure-based paradigm, has already appeared on the horizon – structure-based virtual screening of small-molecule libraries. Advantages include speed, greatly reduced screening costs and the ability to generate multiple backup series for broader patent protection. Future extensions of this approach incorporating computational ADME toxicity, drug-likeness prediction and computational pharmacogenomics are already in development.

Structural databases vs single protein structures – why are databases needed?

Why is it important to have a database of protein structures? Historically, protein structures have been difficult to obtain. Physical methods, such as X-ray crystallography, are slow and earlier modeling technologies were not sufficiently accurate to be definitively useful in drug design. As a result, drug design researchers have typically worked with one protein at a time. Figures 4 and 5 show why it is important for us to stretch our minds a little and to begin to think about (potentially large) groups of related proteins.

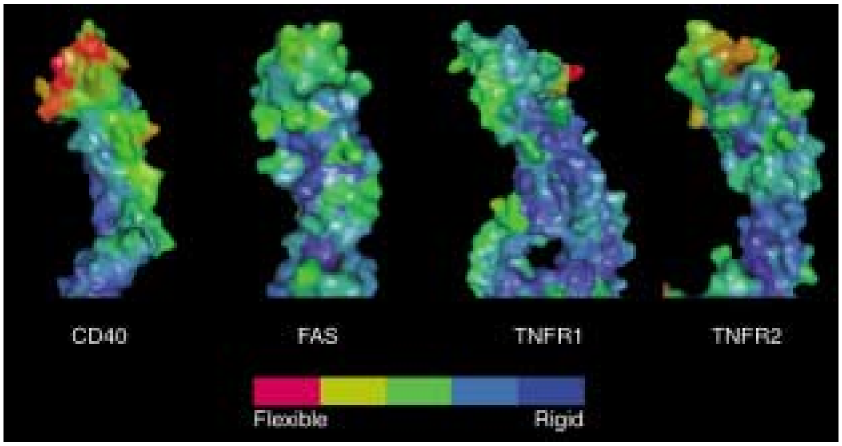


Figure 5. Dynamic surface renderings of TNFR1, TNFR2, CD40 and FAS. Regions colored in blue indicate surface residues that are rigid, and the regions colored red indicate surface residues that are flexible. The rigidity and flexibility measures are based on running a 500 ps molecular dynamics trajectory and analyzing the trajectory to determine the range of movement for the surface-exposed residues.

Comparative analysis of related structures provides insights into the location and nature of potential binding sites. It provides the information required to differentiate related protein family members on the basis of various structural and physical properties. Comparisons can be multidimensional, including properties such as shape, electrostatic charge and dynamic flexibility and many other parameters. In the absence of large-scale, rapid access to protein structure, the time and cost of generating related structures frequently deters comparative analysis. However, this is being changed as a result of the large-scale availability of protein structure through protein structure databases, both those currently available and those that will inevitably follow.

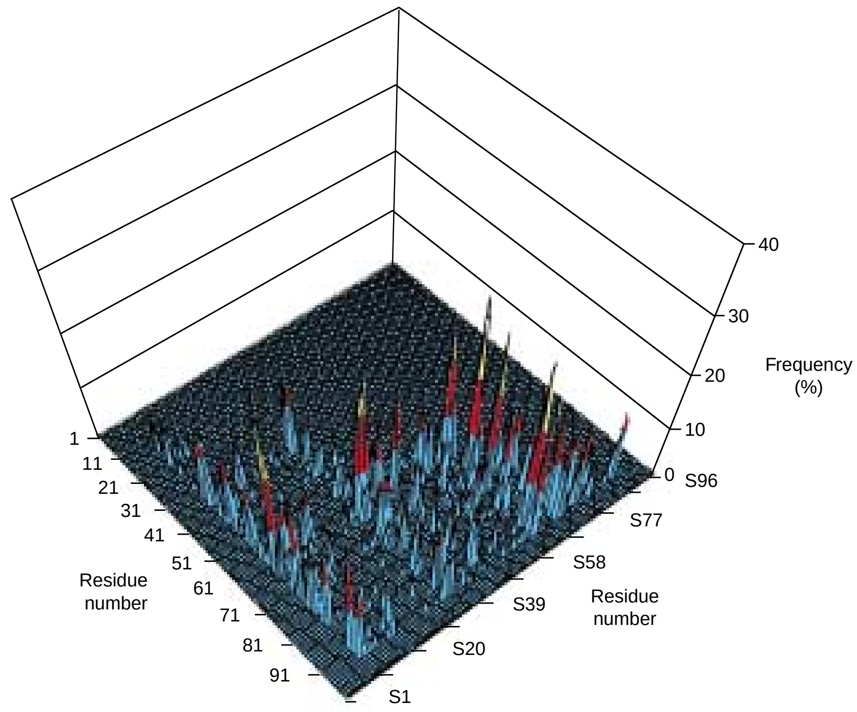
Figure 4 provides an informative example. Tumor necrosis factor receptor 1 (TNFR-1) is an important drug target and blocking it provides remission from rheumatoid arthritis. When Structural Bioinformatics (SB) began a small-molecule drug discovery program for TNFR-1 antagonists, one of the first questions that arose was whether or not it would be possible to create a molecule that could distinguish TNFR-1 from closely related family members – particularly CD40 (another important potential drug target). On the basis of the similarity among protein family members, it was assumed that this would be a difficult task. However, when we analyzed the binding site of TNFR-1 and the corresponding site on CD40, we found that there are radical differences with respect to charge, hydrophobicity and dynamic flexibility. Thus, by analyzing multiple protein structures, it became possible to exploit structural differences among these related proteins to develop specific TNF receptor antagonists. Other members of the TNF receptor family are also shown in Fig. 4, demonstrating differences that can be exploited by the drug designer to develop specific drugs. Therefore, one application of multiple protein structures is to help scientists to build specificity into drugs early on in the design process.

Figure 5 shows that differences in dynamic flexibility also exist among the same subset of TNF receptor family members. Rigid regions of a protein are preferred over flexible areas because flexibility interferes with drug binding. Thus, differences in flexibility observed across a set of related protein family members can be exploited to build in selectivity.

Structural databases in population proteomics and computational pharmacogenomics

Figure 6 shows the paired mutation frequency derived from analysis of a 10 591 unique-structure subset of SB's Variome™ HIV protease structural variant database. Residues have been shown to co-mutate with other residues with defined statistical frequencies.

Paired mutation frequency, mutational stability in enzyme active sites, drug binding regions, or areas of protein–protein interactions and other aggregate structural properties expressed in 3D, as shown in the example of yet another viral protein drug target in Fig. 7 (derived from more than 11,000 unique Variome™ HIV reverse transcriptase structures), can reveal coupled or compensatory structural changes and point to the most effective drug target sites found in the largest fraction of the patient population. This type of information is not available from single structures but only from structural polymorphism databases. SB's Variome™ family of structural variant databases is generated in partnership with Quest Diagnostics (Teterboro, NJ, USA). Variome™ databases permit the practical integration of



Drug Discovery Today

Figure 6. HIV protease paired mutation frequency of a subset of proteins from the Variome^{TM} database. The subset chosen is made up of 10,591 unique structures. This graph shows that residues co-mutate with other residues with defined statistical frequencies.

population proteomics and computational pharmacogenomics into the rational design of so-called 'best-in-class' drugs. This type of information, which will be used increasingly in the design of next-generation pharmaceuticals, scarcely begins to reveal the enormous power of large-scale protein structural databases[30].

In silico (virtual) screening and focused library design – near-perfect structures in an imperfect world

Structure can be used to directly augment screening-based drug discovery, improving efficiencies by three to four orders of magnitude with respect to the number of molecules that need to be synthesized and tested. When a high-quality protein structure is available to be investigated as a drug target, it can be used to rapidly screen databases or libraries of existing chemical compounds to select those compounds that possess an arrangement of chemical groups (charges, hydrophobic groups, hydrogen bonding acceptors or donors) that are compatible with binding to, or mimicking, the targeted active site of the protein. This is a fundamentally different approach to that of X-ray crystallography-based rational drug design, which has been used primarily to optimize lead molecules through examination of co-crystallized drug–protein complexes. A structure-based screening approach, however, attempts only to identify a subset of an existing or a virtual library of compounds with an enhanced probability of binding. The X-ray cystallography-based rational drug design method can be used to pre-screen available or virtual compound libraries that have been optimized for drug-likeness, chemical diversity and, ultimately, predicted ADME/Tox characteristics. Similarly, high-quality protein structure can be used to design focused combinatorial libraries aimed at selected targets. Computational pre-screening can be carried out *in silico* against compounds with unlimited chemical diversity, dramatically reducing the infrastructure requirement, time and cost for synthesis or compound acquisition, accession, storage, retrieval, solution preparation and physical screening by restricting *in vitro* testing to a small set of confirmatory assays and QSAR determinations. Using this type of approach, we have obtained hit rates of 2–24% for the ten targets examined so far, compared with 0.01–0.001% hit rates typically achieved using HTS – a 1000- to 10,000-fold improvement in efficiency. This has been made possible by the availability of high-quality protein structures for the 10 drug targets through computational proteomics.

Function prediction – imperfect structures in a perfect world

Low-quality structures, such as those generated using automated homology modeling methods and threading methodologies, are not particularly useful in structure-based drug discovery or optimization applications. However, they are increasingly finding use in function prediction. Automated homology modeling and threading rely on the identification of common 3D structural elements that can be correlated with a particular enzymatic function. Foreknowledge of the 3D patterns used to search structures for similarity has been a prerequisite of function prediction using low quality protein structure. Currently, a relatively limited number of such patterns are known, discovered by exhaustive manual examination of known enzyme mechanisms. The methods are relatively crude and result in only general predictions about the class of activity associated with a novel sequence.

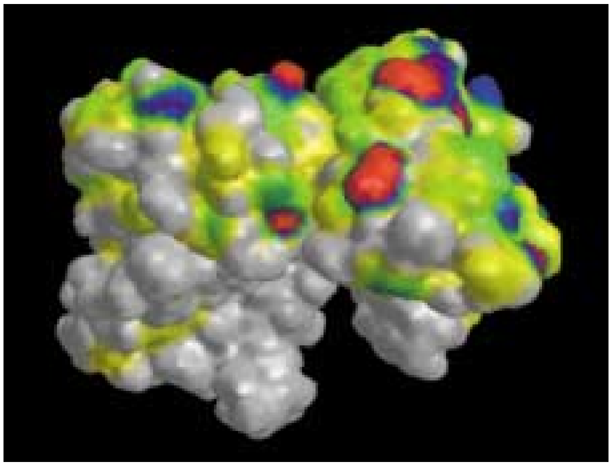


Figure 7. Molecular surface of HIV reverse transcriptase with residue surfaces colored according to frequencies of mutations. The color code for frequencies of residue mutations is as follows: white, <1%; yellow, 1–10%; green, 10–30%; blue, 30–50%; and red, ≥50%.

Protein structural databases might have a significant role in expanding the predictive value of such methodologies. At SB-Moldyn (Cambridge, MA, USA) we developed a structural pattern recognition algorithm, StructureMiner™. This algorithm is now being extended to analyze families of proteins of known function *en masse,* to automatically identify novel 3D functional assemblies that correlate with enzymatic or protein-binding activity. A fundamental requirement of this approach is the broad availability of high-quality 3D protein structural databases. It has the rather significant advantage of not requiring any foreknowledge of the 3D structural ensembles that correlate with protein function. Such an approach should prove particularly useful in identifying novel function among the vast array of expressed protein structures as each new genome is elucidated.

Conclusion

There can be little doubt that computational proteomics will 'make good' its promise of a scientific and commercial revolution of unequaled impact on mankind over the course of the next decade and beyond. The large-scale availability of structural databases will impact all aspects of the life sciences. There is a growing trend that will continue to accelerate towards the increased use of computational methods to model laboratory results before actual physical experimentation. Although the science-fiction vision of achieving ultimate results through pure virtual scientific endeavors remains just a dream, it is clear that the greatest advances in the near future will be made through the integration of computational methods with physical data from X-ray crystallography, mass spectroscopy and NMR, along with biological and chemical experimental results.

Knowledge of protein structures will have a major role in increasing the efficiency of drug discovery. Structure-based computational pre-screening of real and virtual small-molecule libraries, along with the structure-guided generation of focused combinatorial libraries, will increase the speed of discovery and optimization of new drug molecules. The large-scale use of structural variant databases will enable drug companies to design drugs that function in the largest fraction of the patient population – a key characteristic of 'best-in-class' drugs. The broad use of drug target databases that incorporate not only structures of interesting targets but also of closely related protein family members will permit these companies to design drugs that have maximum selectivity and, therefore, minimum side effects.

The molecular biology and biochemical research fields will be radically transformed by the broad availability of 3D protein structural information. In the past, scientists typically conducted site-directed mutagenesis, chemical modification, assay design and similar activities on a more or less empirical basis but, in the future, all such experiments will be rationally designed with full knowledge of the protein structure at hand. Just as an organic or medicinal chemist would not consider conducting a synthetic reaction without first carefully considering the underlying structural issues to predict the probable outcomes and reaction products, molecular biologists and biochemists will no longer think of proceeding with their research in the absence of full structural information. This will be made possible through the broad availability of protein structure on a large scale provided by protein structural databases and increasingly informative and powerful 3D and 4D protein structural data mining tools.

Acknowledgement

Commissioned for, and first published in, *Trends in Biotechnology* (July 2001).

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

3 Norwell, J.C. and Machalek, A.Z. (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nat. Struct. Biol.* 7, 931

4 Terwilliger, T.C. (2000) Structural genomics in North America. *Nat. Struct. Biol.* 7, 935–939

5 Yokoyama, S. *et al.* (2000) Structural genomics projects in Japan. *Nat. Struct. Biol.* 7, 943–945

6 Heinemann, U. (2000) Structural genomics in Europe: slow start, strong finish? *Nat. Struct. Biol.* 7, 940–942

7 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242

8 Panchenko, A. *et al.* (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins Struct. Funct. Genet.* 37(Suppl. 3), 133–140

9 Padilla, C.E. and Karlov, V.I. (2000) Method for generating a refined structural model of a molecule. US Patent No. 6,125,235

10 Chun, H.M. *et al.* (2000) MBO(N)D: a multibody method for long-time molecular dynamics simulations. *J. Comp. Chem*. 21, 1–26

11 Borchert, T.V. *et al.* (1993) The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: the correct modeling of an eight-residue loop. *Structure* 1, 205–213

12 Eisenmenger, F. *et al.* (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* 231, 849–860

13 Vasquez, M. (1996) Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* 6, 217–221

14 Balaji, V.N. and Singh, U.C. (1997) Method of rational design based on *ab initio* computer simulation of conformational features of peptides. US Patent No. 5,612,895

15 Dudek, M.J. *et al.* (1998) Protein structure prediction using a combination of sequence homology and global energy minimization: II. Energy functions. *J. Comp. Chem*. 19, 548–573

16 Ramnarayan, K. *et al.* Protein structural bioinformatics. *Pharmacogenomics* (in press)

17 Chou, P.Y. and Fasman, G.D. (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* 13, 211–222

18 Garnier, J. *et al.* (1978) Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120, 97–120

19 Rost, B. *et al.* (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 323, 584–599

20 Geourjon, C. *et al.* (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus predictions from multiple alignments. *Comput. Appl. Biosci.* 11, 681–684

21 Frishman, D. *et al.* (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins Struct. Funct. Genet*. 27, 329–335

22 Cuff, J.A. *et al.* (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins Struct. Funct. Genet.* 34, 508–519

23 Chandonia, J-M. *et al.* (1999) New methods for accurate prediction of protein secondary structure. *Proteins Struct. Funct. Genet.* 35, 293–306

24 Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202

25 Petersen, T.N. *et al.* (2000) Protein secondary structure prediction at 80% accuracy. *Proteins Struct. Funct. Genet.* 41, 17–20

26 Bohr, J. *et al.* (1997) Protein folding and wring resonances. *Biophys. Chem*. 63, 97–106

27 Skolick, J. and Fetrow, J.S. (2000) From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol.* 18, 34–39

28 Jones, D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature* 358, 86–89

29 Lemer, C.M. *et al.* (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins Struct. Funct. Genet.* 23, 337–355

30 Ramnarayan, K. *et al.* (2000) US Patent Application Serial No. 09/438,566