Biotechnology, Drug Discovery, and Biomedical Research *In Silico*: The State of the Art

Dirksen E. Bussiere
Information Management and Technology
and Department of Structural Biology
Abbott Laboratories
Dept. 42T, Bldg. AP10-L7
100 Abbott Park Road
Abbott Park, Illinois 60064-3500
bussierd@asok.pprd.abbott.com

INTRODUCTION

During the last two decades, new techniques and advancements have added layers of complexity to the research done in biotechnology, drug discovery, and biomedical research. Many of the techniques developed have been computational in nature: in effect, taking their particular disciplines *in silico* (that is, to be executed within the computer). Often, these techniques are lumped under the term of 'biomedical computing'. To visualize these advancements, compare the 'traditional' drug discovery cycle (Figure 1) used by many companies until the early-1980's with the current biological target-based drug discovery cycle (Figure 2) used by most pharmas and biotech concerns today. In the 'traditional' cycle, an initial lead compound was found by isolating a molecule with a certain biological activity, perhaps by ethanobotany or by fortuitous happenstance (as in the discovery of penicillin); modifications of this lead compound were planned using clues provided by a crude analysis of structure-activity relationships (SAR) or by traditional medicinal chemistry techniques. The new, modified compounds were then synthesized and retested. This cycle continued until the particular biological activity of the compound was maximized [1]. This cycle, while successful, was not rapid. It often took 5-6 years to bring a drug to the preclinical phase.

Contrast this 'traditional' cycle to the current biological target-based cycle (also known as structure-based drug design; Figure 2)[2]. Now, recognizing that many drugs are antagonists (inhibitors) of macromolecules (most often an enzyme protein), a ' biological target' molecule is chosen before any drug discovery project is begun. The biological target is a macromolecule which is crucial for the biological activity or process which is to be inhibited. In some cases, biological target selection is simply a matter of common sense and examination of results from basic research. For example, human immunodeficiency virus-1 (HIV) is expressed as a single polypeptide within an infected host cell. This polypeptide is then processed by an virally-encoded protease; the processed proteins are then packaged and the virus explodes from within the infected cell. It was therefore correctly surmised that HIV protease was critical for virus maturation and was an important biological target for drug discovery and development. This has led to several highly effective therapeutics for HIV[3]. It is not always that simple, however, especially when the biological activity to be inhibited is not parasitic in nature (as in the case of a viral infection) or when the number of possible targets is enormous. In these cases, the computational field of bioinformatics plays a critical role by providing methodology for scanning the current genomic databases for an optimal target or targets and for predicting the activities of the as-of-yet uncharacterized genes and proteins.

Assuming that one is able to select a target, several technologies come into play. First, the gene of the target of interest is cloned and the protein or macromolecule is expressed and purified. The initial lead compound is then discovered by a variety of techniques such as high-throughput screening, where hundreds of thousands of compounds are examined *en masse* for binding to the purified target (Figure 2). Often, in a concurrent effort, the three-dimensional structure of the target macromolecule will be determined using nuclear magnetic resonance (NMR) or X-ray crystallography, or the structure can be modeled using molecular modeling techniques. The structural methods are extremely dependent on computational techniques, as will be discussed in a following section. Once the structure of the target macromolecule has been determined or modeled, and a lead compound has been isolated, the structure of the target-compound complex can be determined using the same techniques. These target-compound structures can then be examined using computational chemistry techniques and possible modifications to the compound can be determined. Finally,

all of this data is collated and is used in designing the next series of compounds, which are then synthesized. This cycle is repeated until a compound is sufficiently potent (able to inhibit the biological target at extremely low, typically picomolar, concentrations), at which point it is sent to preclinical (animal testing) and clinical (human) testing[2]. In the current discovery cycle, an average time to reach preclinical investigation is three years.

While the following example was obviously slanted towards small-molecule drug discovery, the computational tools play a similar role in biotechnology and biomedical research. The following special session will present cutting-edge research in the area of biomedical computing. A brief review and introduction to each section that will be covered follows.

COMPUTATIONAL CHEMISTRY

Perhaps one of the most ambitious undertakings is the ongoing attempt to model chemical systems *in silico*. The field began quite humbly in the late 1960's with a simulation of a 'box' of 216 waters on one of the fastest computers of the time[4]. Now, such molecular dynamics simulations often involve proteins comprised of tens of thousands of atoms 'immersed' in boxes or spheres of thousands of water molecules and are run a a desktop machine or local server[5]. Similarly, the toolbox of the computational chemist now includes techniques for estimating the strength of binding of a small molecule to a biological macromolecule, the ability to do complex electrostatic calculations on macromolecules, and modeling techniques which allow the modeling of novel macromolecules from other macromolecule structures of nominal homology; there are many more techniques, far too numerous to list here, as the field is quite dynamic with new techniques being developed each day[6].

Of primary concern to any organization involved in drug discovery, be it a pharma, biotech, or an academic research effort, is the ability to develop a molecule capable of inhibiting a biological target: the elusive lead compound. Our first speaker, Dr. Regine Bohacek of Ariad Pharmaceuticals, will present her research on automated lead generation and the development of the program GROWMOL in her talk entitled "Exploring the Universe of Organic Molecules for Potent New Drugs". GROWMOL is capable of generating organic molecules complementary to the binding site of a macromolecule assuming that the three-dimensional structure of the macromolecule is known[7,8]. It has been successful in generating inhibitors for a variety of macromolecules and promises to add another wonderful tool to the computational chemist's toolbox. Dr. Bohacek is a research scientist at Ariad Pharmaceuticals.

PROTEIN ENGINEERING

The next area to be presented is-on the computational level-an ambitious extension of computational chemistry and molecular modeling. Of course, there have been significant experimental contributions as well. In the early 1980's, researchers began to attempt to rationally engineer protein molecules in such a way as to modify existing structure, stabilize existing proteins against the stresses of heat or salt, add new activities or to modify existing activities, and even to design novel proteins *de novo*[9]. The successes in this area have been quite impressive, especially in the last five years. As one might expect, molecular modeling has always played a pivotal role in these experiments. Often, however, its contribution has been qualitative in nature by allowing researchers to visualize the effect of point mutations within a protein structure. Recently, researchers have been attacking the problem of *de novo* protein engineering with more quantitative approaches which demand, as one might imagine, a significant amount of computation. *De novo* protein design is essentially an attempt to solve the inverse-folding problem: one is attempting to predict the structure of a protein given only its sequence or, vice-versa, given a particular fold or protein structure, to predict the sequences which can adopt that structure. The complexity of this can be seen in the recent work of Dahiyat and Mayo (Stanford University), who were able to predict a non-native sequence for the 23-residue zinc finger fold of Zif268; this problem required a search of approximately 10^{62} different representations of the protein[10].

Our speaker on computational protein engineering, Dr. Homme Hellinga (Duke University Dept. of Biochemistry), was the first scientist to successfully introduce a new binding site *de novo* into an existing protein structure, in this case the introduction a copper-binding site into *E. coli* thioredoxin[11]. His work in protein engineering has been aided by a program which he has developed called DEZYMER[12]. Over the past several years, he has extended his research and his successes in the area of protein engineering to include the introduction of a FeS_4 cubiodal cluster into thioredoxin [13] and the introduction an extraneous iron superoxide dismutase active site into thioredoxin [14]. The title of his talk will be "Structure-Based Design of Protein Function". Dr. Hellinga is an associate professor at Duke University.

STRUCTURAL BIOLOGY

In the last decade, structural biology has transformed the face of biomedical research. Due to the improvement of equipment and experimental methodology, the advent of cloning, and the availability of rapid, inexpensive computers, the solving of macromolecular structures by

NMR and X-ray crystallography has become commonplace. These techniques are allowing us to view the molecules that make up a living organism at an extraordinary resolution, sometimes better than 1 Angstrom (atomic level resolution). However, this is not to say that it is easy. To solve a novel macromolecular structure (be it protein, DNA, RNA, etc.) by NMR involves specific isotopic labeling of that molecule, collection of data, and assignment of resonances within that data, followed by initial structure generation and refinement of that structure. This process can take several months to several years to complete. Likewise, to solve a novel macromolecular structure by X-ray crystallography involves crystallization of the macromolecule in question, introduction of specific heavy-atoms into the protein crystal, collection of data in-house or at a synchrotron, location of the heavy-atom sites, and initial generation of electron-density maps; this is followed by initial structure generation and refinement. As in the case of NMR, structure solution can take up to several years to complete, with the average being approximately 1-2 years.

While this time is not prohibitive in a typical research effort where attention is directed towards a single protein, current cutting-edge research is demanding a more rapid approach. In conjunction with the massive genomic sequencing efforts underway, structural biologists are initiating structural genomic efforts where representative proteins from each genome are solved and used-in conjunction with molecular modeling-to predict the structure (or at least the fold) of homologous proteins within the genome in question and in other genomes as well [15]. Our speaker in this area, Dr. Thomas Terwilliger (Los Alamos National Laboratory) is directing such a structural genomics effort aimed at investigating the structure of the proteins coded within the genome of *Pyrobaculum aerophilum*, a hypothermic archebacteria[16]. To aid in this research, as well as to help speed crystal structure solution in general, Dr. Terwilliger has written a wonderful program called SOLVE which rapidly completes analysis of the diffraction data and generation of interpretable electron-density maps[17]; Dr. Terwilliger and his collaborator, Dr. Joel Berendzen were honored with an R&D 100 award for SOLVE in 1998. With the assistance of SOLVE, Dr. Terwilliger and his team have already solved (no pun intended) three protein structures from the genome in the time it takes a typical lab to solve one structure[18]. Dr. Terwilliger's talk is entitled "Structural Genomics and Automatic Methods for Macromolecular X-ray Structure Determination". Dr. Terwilliger is a senior technical staff member at Los Alamos National Laboratory.

BIOINFORMATICS/COMPUTATIONAL GENOMICS

The last five years has seen an unforeseen amount of activity and success in the area of genomic sequencing, where researchers determine the actual DNA sequence of a genome of an entire organism. At last count over 269 genomes are currently being sequenced or have been completed, spread between viruses (more than 186), procaryotes (65), and eucaryotes (18). It is anticipated that the sequencing of the human genome will be completed by the year 2002. Of course, this data is worth little without the tools to interpret and mine it. This is a daunting task and program suites have been written to identify putative genes (also known as Open Reading Frames or ORFs), promoter regions, and other significant landmarks within this sequence data[19]. This, of course, is just the tip of the iceberg. Two of the talks within the session will focus on what can be done to mine, interpret, and exploit this gold-mine of data. The interpretation of the data from these projects is of significant interest to pharma, biotechnology concerns, and academic biomedical research, as each ORF might represent a significant biological target.

Our first speaker in this area, Dr. Mary Donlan (Molecular Simulations, Inc.) will describe her company's efforts to utilize novel algorithms and the large bulk of available protein structural data to predict and model proteins predicted to exist (again as ORFs) in genomic databases. This would allow the description and evaluation of the secondary/tertiary fold and active site of all identified ORFs within a genome, with an eye towards discovering suitable biological targets for structure-based drug design. Dr. Donlan is senior product manager at MSI in charge of their structural biology and bioinformatics efforts. Previously she was a scientist at Glaxo Research Institute. Her talk is titled "3D Protein Databases Integrated with Genomic and Chemical Data".

Dr. Inna Dubchak (Lawrence Berkeley National Laboratory) will describe her work on developing neural networks capable of predicting the three-dimensional protein fold of a putative gene, discriminating between introns (non-coding regions) and exons (coding regions) within putative genes (ORFs), and identification of putative RNA genes in genomic data[20]. This ambitious approach would allow a significant and comprehensive mining of any genomic data. Her talk is titled "Global Description of Amino Acid and Nucleotide Sequences: Application to Predicting Protein Folds, Intron-Exon Discrimination, and RNA Gene Identification". Dr. Dubchak is a computer scientist at the Center for Bioinformatics and Computational Genomics at Lawrence Berkeley National Labs.

Our final speaker will bridge from bioinformatics and computational genomics to mathematical biology. A living

organism, be it a bacteria or a human, is far more than a collection of enzymes, each catalyzing a single reaction. Rather it is a dynamic system with significant synergy existing within a cell and with significant cross-talk between individual components. Dr. Frank Tobin (SmithKline Beecham Pharmaceuticals) and his group have been investigating modeling these genetic systems mathematically (see Dr. Tobin's summary for a comprehensive reference list). This is of significant interest to many members of the scientific community because if an accurate model of such a system can be constructed, one would be able to predict the effect of perturbations (such as a gene knockout or enzyme inhibition) on the system under investigation without the need to do lengthy and expensive *in vivo* experiments. Dr. Tobin will describe his group's efforts and methodology in his talk titled "Reconstructing Gene Regulatory Networks from Gene Expression Data". Dr. Tobin is associate director of mathematical biology at SKB.

CONCLUSION

Obviously, it is impossible to present all the current work being done in the fields of biomedical computing, especially in a one-day session. However, this session should allow one to get exposure to, and a sense of, current topics of interest to the research community in each area. Each speaker and their research is representative of the type of work ongoing in that area. It is the hope of the organizer that this session will allow the audience to gain a perspective on the opportunities and challenges that each area holds. And, as is the case of most cutting-edge research, the lines between individual areas often blur and cross-disciplinary work ensues. The opportunities for productive research and program development are too numerous to list.

REFERENCES

[1] P.A. Case and M.J. Balick. Scientific American **270**, 82-87, 1994.

[2] J. Greer, J.W. Erickson, J.J. Baldwin, and M.D. Varney. J. Med. Chem. **35**, 1035, 1994.

[3] A. Wlodawer and J. Vondrasek. Ann. Rev. Biophys. Biomol. Struct. **27**, 249-284, 1998.

[4] M. Gerstein and M. Levitt. Scientific American, **279**, 100-105, 1998

[5] R. Elber. Curr Opin. Struct. Biol. **6**, 131-135, 1996.

[6] T.J. Marrone, J.M. Briggs, and J.A. McCammon. Ann. Rev. Pharmacol. Toxicol. 37, 71-90, 1997.

[7] R.S. Bohacek and C. McMartin. J. Am. Chem. Soc. **116**, 5560-5571, 1994.

[8] D.J. Rich, R.S. Bohacek, N.A. Dales, P. Glunz, and A.S. Ripka. In *Actualites de chimie therapeutic*, Elsevier: Amsterdam, 101-111, 1996.

[9] W.H. Rastetter. Trends Biotech. **1**, 80-84, 1983.

[10] B.I. Dahiyat and S.L. Mayo. Science **278**, 82-87, 1997.

[11] H.W. Hellinga, J.P. Caradonna, F.M. Richards. J. Mol. Biol. **222**, 787-803, 1991.

[12] H.W. Hellinga and F.M. Richards. J. Mol. Biol. **222**, 763-785.

[13] C.D. Coldren, H.W. Hellinga, and J.P. Caradonna. Proc. Natl. Acad. Sci. USA **94**, 6635-6640, 1997.

[14] A.L. Pinto, H.W. Hellinga, and J.P. Caradonna. Proc. Natl. Acad. Sci USA **94**, 5562-5567.

[15] A. Sali. Nature Struct. Biol. **5**, 1029-1032, 1998.

[16] T.C. Terwilliger, G. Waldo, T.S. Peat, J.M. Newman, K. Chu, and J. Berendzen. Protein Sci. **7**, 1851-1856, 1998.

[17] T.C. Terwilliger and J. Berendzen. In preparation.

[18] T.S. Peat, J. Newman, G.S. Waldo, J. Berendzen, and T.C. Terwilliger. Structure **6**, 1207-1214, 1998.

[19] D. Benton Trends Biotechnol. **14**, 261-272, 1996.

[20] I. Dubchak, I. Muchnik, and S.H. Kim. Microb. Comp. Genomics **3**, 171-175, 1998.

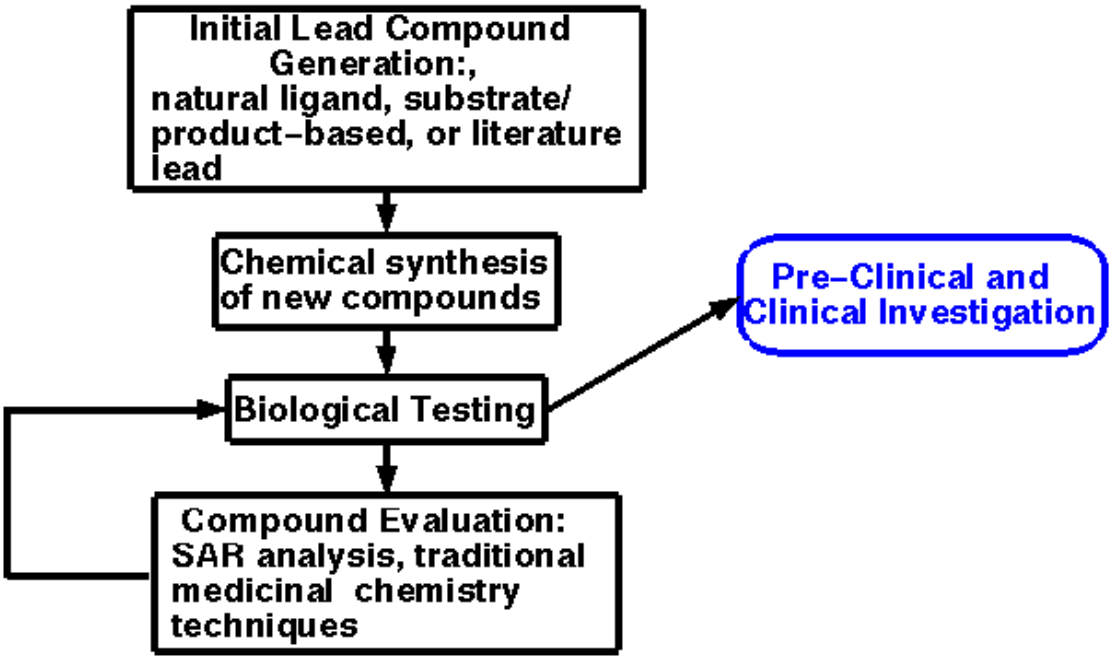

Figure 1
Drug-Discovery and Development pre-1980

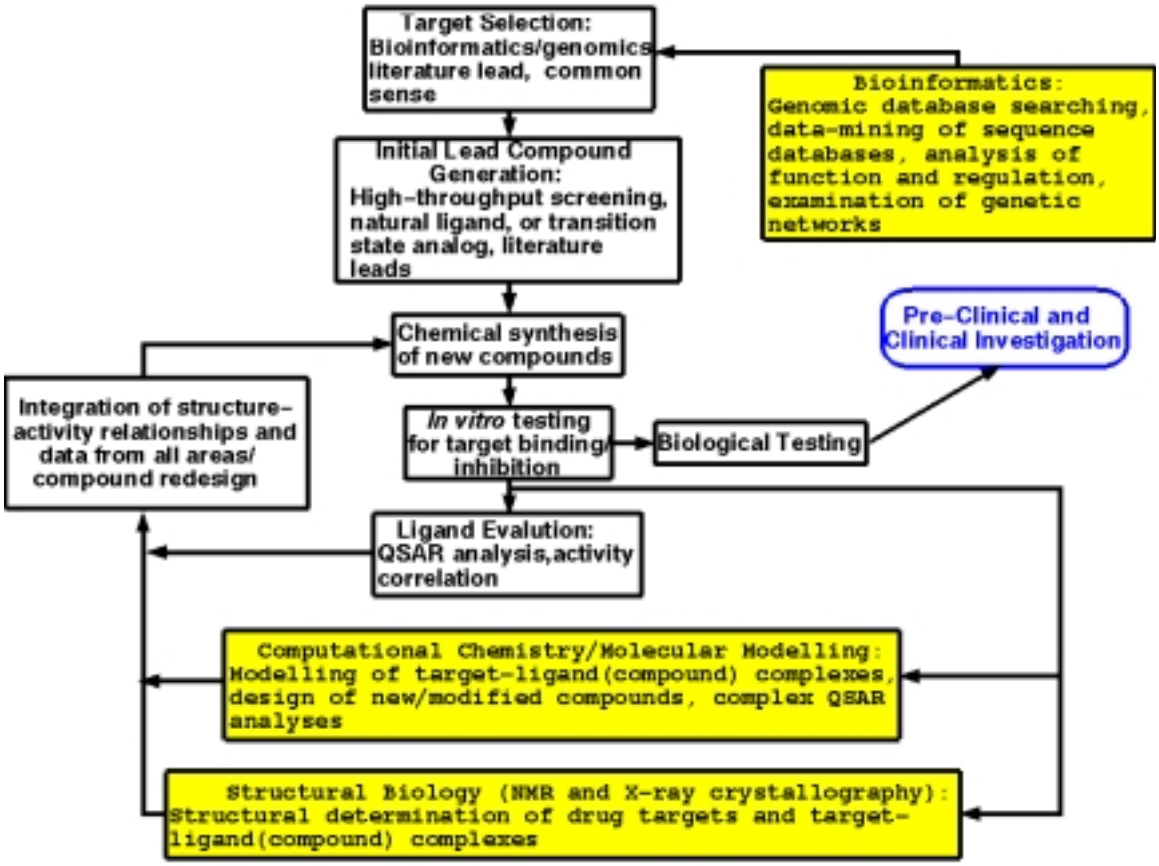

Figure 2
Drug Discovery and Development Today:
the Structure-Based Drug Design or 'Rational' Drug Design Approach