Drug Design In Post Genomic Era

Anuradha, C. M*
Department of Biotechnology,
Sri Krishnadevaraya University, Anantapur-515 003. AP, India.

Abstract: The drug development process creates a crystallized version of the drug molecule which is also patented. After the initial molecule is patented, the process of further research, developing the drug, and conducting clinical trials takes about eight years. Applying rapid assay techniques to the compounds produced using combinatorial chemical technology allows the creation of large chemical populations of molecules which after annotation will be placed at ligand database. A pharmaceutical company would like to test on as broad a population as possible while still being able to detect any treatment effect. Another area of interest is the choice of endpoints.

Index Terms – Drug Design, Genome, Lead molecule.

Introduction:

In genomic era Bioinformatics and Chemiinformatics are playing a key role in pharmaceutical industry to design new drug targets from genomic data at very faster rate. We have to note how genomics has dramatically altered the way of drug discovery. Disease causing genes are identified using the tools of genomics and proteomics. The process of designing a new drug using bioinformatics tools has been of great help in identifying target disease, interesting lead compounds, and by docking studies finding the effective interaction between drug and the target compound. Proteomics has unique and significant advantages as an important complement to genomics approach specifically for Target/marker identification and Target validation/toxicology, which are very crucial in making lead molecule as drug. The drug industry is very much where a winner-takes-all game. The first company to patent a drug for a particular therapy gets exclusive rights in its use (for 20 years), while the runner-up is given nothing. The winning edge could mean the difference between a billion dollars of revenue and of years of work lost.

Drug discovery is a time consuming and expensive process. It is becoming increasingly difficult to find new compounds that will lead to new drugs. The top twenty pharmaceutical companies spent more than $1600 billion on research and development during last two years. The current method of identifying new drugs focuses on finding biologically active candidates from different sources. Combinatorial chemistry has received much attention in this respect. The drug development process creates a crystallized version of the drug molecule which is also patented. Determining the crystal structure and its possible polymorphs is a big challenge. The next step is formulating the drug. Here the focus is on obtaining a method of delivering the drug in a means most comfortable for the patient while considering cost and logistic issues. These issues include the required release rate, costs of formulating the mechanism (powder, liquid, etc), and the form of the crystal structure (if the molecule indeed crystallizes). Once the final formulation of the drug is developed, clinical trials can begin. After the initial molecule is patented, the process of further research, developing the drug, and conducting clinical trials takes about eight years. Since patents only last 20 years, the development process leaves only 12 years for the company to recoup the research and development costs. The question now is: How is a potential drug identified? The answer currently lies in the field of combinatorial chemistry.

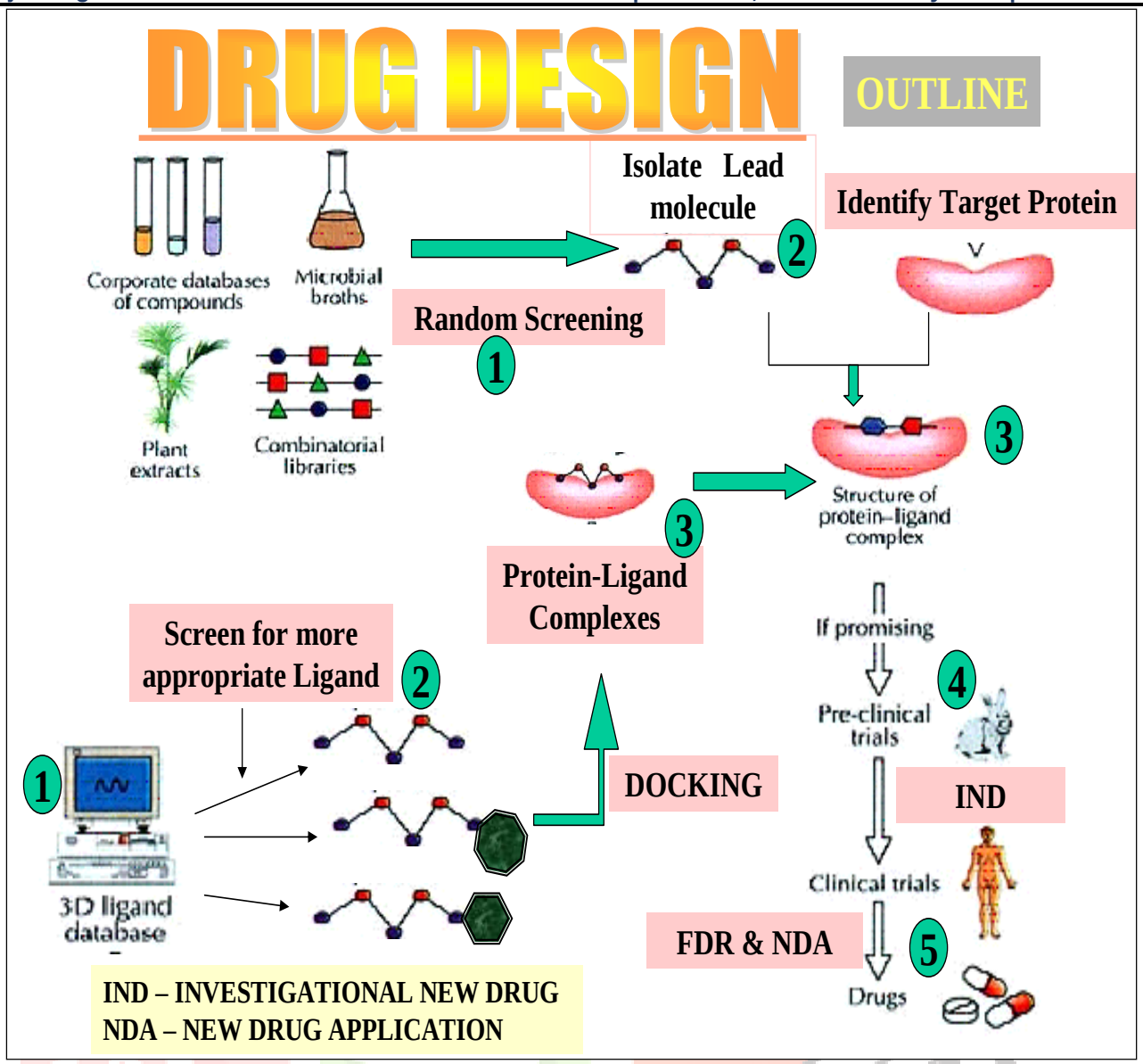


Fig.1 Outlines of structure based and computational drug design.

Combinational Chemistry

The current trend in the search for potential molecules concentrates on the preparation of "chemical libraries" or ‘’ligand database”. A chemical library is a collection of differing molecules created intentionally in a systematic fashion. The molecules can be prepared synthetically or biosynthetically and screened for biological activity and made available to public through internet. Combinatorial chemistry is a synthetic strategy which leads to large chemical libraries and generally defined as ``the systematic and repetitive, covalent connection of a set of different `building blocks' of varying structures to each other to yield a large array of diverse molecular entities." With the development of high throughput, automated techniques, it is now possible to screen hundreds of thousands of individual compounds per year, per drug target using *in silico* technology. Applying rapid assay techniques to the compounds produced using combinatorial chemical technology allows the creation of large chemical populations of molecules which after annotation will be placed at ligand database. To begin the construction of a library, an assortment of small, reactive molecules (chemical building blocks), is essential. There are a number of criteria that are considered when deciding which blocks to use. Building blocks should exhibit a variety of physiochemical properties, functionality, charge, conformation, etc. Each building block undergoes a series of reliable, high yielding reactions to produce a population of related products. The library should be a stable population of molecules with low molecular weight that are neither reactive nor toxic. It is important that the library contains members that are capable of interacting with the biological target of interest. There are two types of strategies within the combinatorial chemistry framework. Broad-based (random) screening creates a vast library containing very diverse structures in the hopes of identifying a ligand of significant affinity for the target. This strategy incorporates less structure in the methodology, with fewer pre-conceived notions about the structure of active molecules. However, once a lead is available, the drug discovery focus changes to a chemical analogy

(directed screening) or optimization strategy. Here the idea is to create a diverse population that closely resembles the original molecule to optimize biological potency. In this case the library is smaller, the structure is limited in its diversity, and the methods involved are more specific and defined.

Clinical Trials

Once a drug is discovered and formulated, clinical trials begin. It is in this area of the drug development process where statistics plays a large role. As a result of the thalidomide tragedy, in 1962 the Food Drug and Cosmetic Act was amended which required evidence of effectiveness before a drug could be marketed. Statistics moved to the forefront as statisticians participated in the design, implementation, and analysis of clinical trials. Currently, emphasis has moved to the efficiency and timeliness of the drug review process as both the public and industry demand prompt reviews and access to experimental drugs. There are also some statistical issues which continue to be discussed as they have not yet been resolved - multiplicity, non-compliance and data integrity are examples. The regulatory process demands that a strict methodology be used in the design, implementation, and analysis of clinical trials. When reviewing a new drug application, a regulatory board is interested in a number of features of the trial. The population analyzed is important because it affects the generalizability of the findings. If the drug was tested on a group of men between the ages of 25-40, any drug effect found has not been proven for women above the age of 65, for example. A narrow population tested makes it easier to find an effect, but restricts the applicability of the findings. On the other hand, too broad a population focus and an effect may not be found as it is diluted in the variety of the subjects tested. A pharmaceutical company would like to test on as broad a population as possible while still being able to detect any treatment effect. Another area of interest is the choice of endpoints. To determine if a treatment is having an effect, some response must be measured. One response is designated the primary response variable. Frequently, however, the outcome of real interest is not measurable. There are two main reasons why this may occur: the clinical response takes a long time to occur on average, or the response of main interest is difficult to observe or measure. Examples of the first include measuring CD4 counts or tumor shrinkage instead of death as the outcome. An example of the latter is measuring plaque buildup on the carotid artery (in the neck) when the response of interest is the buildup on the coronary arteries, since measuring within the heart is an invasive and dangerous procedure. It is hoped that the effects are comparable between the measured response and the actual one, but the connection between the clinical outcome and the surrogate endpoint must be established. Usually in a study using surrogate outcomes, more than one outcome is used in the examination of a treatment effect and then a means of coping with more than one response must be found (the normal procedure is to consider one outcome the primary endpoint and all others as secondary outcomes). Sample size is a very important determinant to finding an effect. Before beginning a study, the number of subjects needed to detect a treatment effect is calculated. Without enough subjects, no statements can be made concerning the effect. The sample size depends on a number of factors: the variability in the population, the desired error rates, and the magnitude of the effect (the detectable difference between the means of the treatment and control groups, say). The greater the number of subjects, the better the chances of finding an effect. However, each additional subject requires more resources. These two considerations must be balanced when considering the required sample size. A key aspect to the validity of any results is the method of randomization. The concept of randomization is an integral part of the statistical process as it allows the experimenter to control (in the long run, for large samples) the known and unknown effects that will confound the study - that is, those variables that we do not explicitly control for in the design and analysis that may interfere with the results. Being such an important factor in making statistical statements, the review board is concerned with the methods used to randomize the subjects to the treatments. The placebo effect is a well-documented phenomenon whereby people feel better when taking a treatment, though they are only receiving a sugar pill. To control for this effect, because knowing the treatment may affect the responses of the patient, the subject should be unaware of which treatment they are receiving. This is single blinding. An additional safeguard is double blinding, where those evaluating the treatment effects do not know what treatment the subject is receiving. Finally, the review board investigates whether the actual statistical analysis is equivalent to the statistical analysis plan. In a clinical trial protocol, the pharmaceutical company is required to describe the planned methods of analysis, in specific detail. Deviations from the plan are not encouraged by the review board as it raises the question, ``Why?'' Any changes in the analysis approach must be justified in the drug application to the review board.

Genomics and Proteomics in Drug Discovery

One has to note how genomics has dramatically altered the way of drug discovery. Now the human genome sequence is known available to public (at NCBI) and we need to "mine" the sequence data ("data mining") to find out what it means results in offering very rich prospect of finding better drug targets.

What needs to be done is to annotate the genome sequences and search for genes responsible for disease development. After annotating the specific DNA sequence perform translate programs to get protein sequence. Once protein primary structure is established perform the data base search for similarity proteins in other organisms (NCBI-BLAST). Find out the structural coordinates of the similarity reported proteins from PDB and make 3D model of hypothetical protein. Once the 3 D structure targeted protein is build it can be used for ligand interaction studies with dock or prodock to find out the possible best drug molecules.

However, there may not be a good correlation between gene expression and protein expression as most disease processes and treatments are manifest at the protein level. It is believed that gene-based expression analysis alone will be totally inadequate for drug discovery. Proteomics has unique and significant advantages as an important complement to a genomics approach.

Target/marker identification - This application of proteomics provides a protein profile of a cell, tissue and/or bodily fluids that can be used to compare a healthy with a diseased state for protein differences in the search for drugs or drug targets.

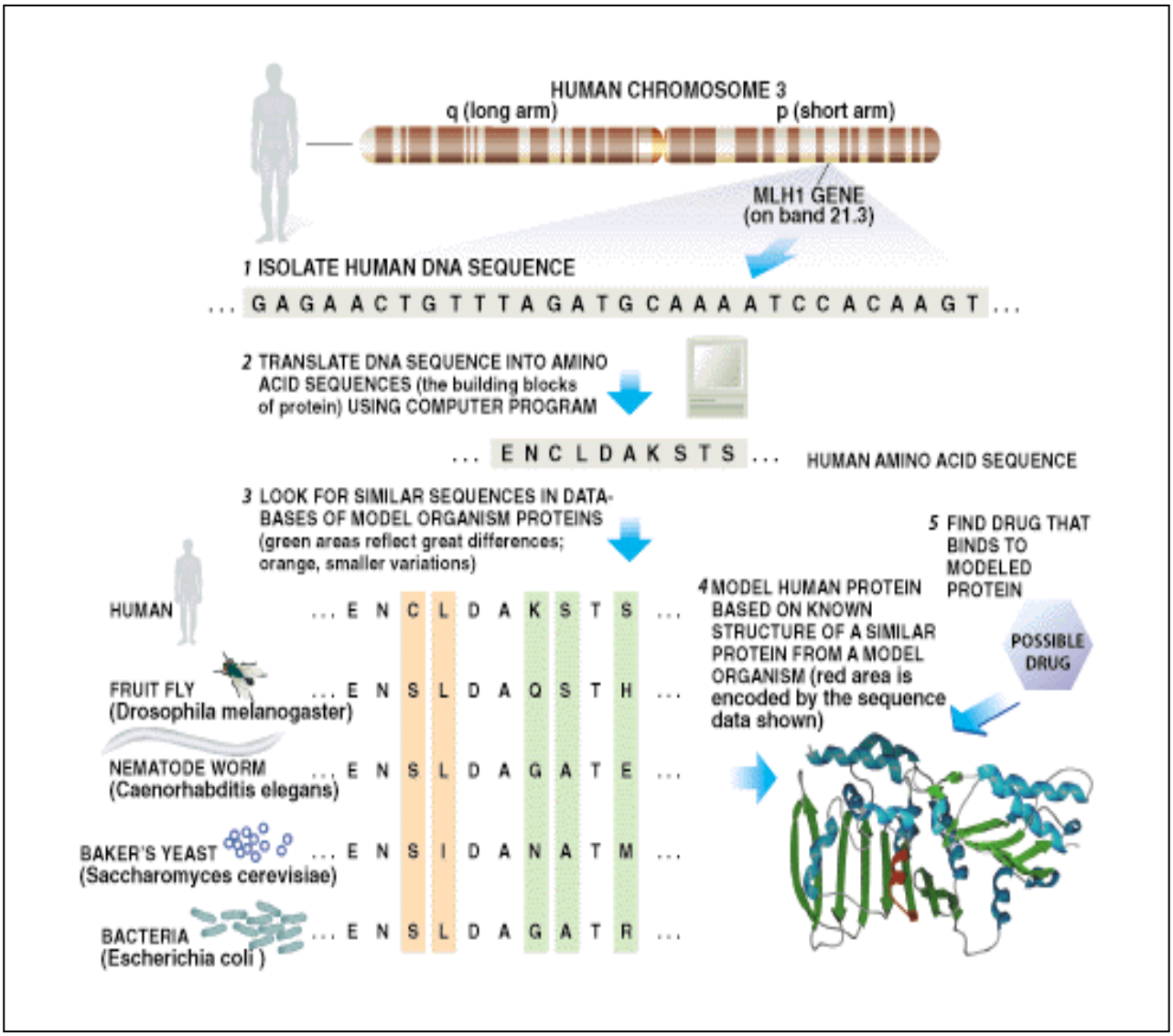


Fig. 2 Outlines of drug design from Genome sequences

Target validation/toxicology - Proteomics can be applied as an assay procedure for the potential utility of drug candidates. This can be achieved by a comparative analysis of reference protein profiles from normal or diseased states with profiles after drug treatment. Proteomics technology can also be integrated with combinatorial chemistry to evaluate comparative structure-activity relationships of drug analogs. A variation of target validation is to study the toxicity of drugs by proteomics. A comparison of the protein profiles from normal tissue or tissue treated with the known toxic agent might give an indication of the drug's toxic activity. An example of this was the study of the toxic activity of cyclosporine A (CsA) in kidney. 2D gel profiles of rat kidney proteins with or without CsA treatment were compared. One of the protein spots was identified as calbindin which is found in the tubules of the kidney and is involved in calcium-binding and transport. There is good evidence that the toxic effects of CsA are linked to the decrease of calbindin. 2D gel analysis of kidney tissue gave new insights into the side effects of CsA. And as with target validation, a database of the protein profiles after treatment with known toxic agents can be used as a reference for comparative purposes when investigating the side effects of new drugs. Proteomics technology is advancing by leaps and bounds, but the hurdles are greater than anything molecular biology has yet to overcome. DNA can be amplified; proteins cannot. DNA is a linear code defined in the 1950's; proteins fold in baffling ways and interacts unpredictably. DNA is basically static; proteins change in various ways even in an individual cell over a short period of time. But the protein realm is essential to progress in biology. It is the whole basis of fields like diagnostics and pharmaceuticals.

Structure based Drug design

In any drug design strategy, the essence of the problem is to find protein key to match a given enzyme lock. Structure-based experiments to this end have attempted to work backwards from the physical shape of the enzyme (lock) to a protein or other molecule which interacts strongly with it's active site (key). The first step in this problem is to develop tools to evaluate the 3D shape of the enzyme, which can also be used to examine/select possible drugs. Several classes of drugs exist, as discussed elsewhere. One of particular interest to structure based development initiatives (financially and as a focus for proteomics) are known as Small Molecule Inhibitors (SMIs). SMIs are small proteins, usually only one or two subunits of less than 150 amino acids total size. They usually attempt to inhibit biological pathways via the most simple mechanism; extremely efficient binding at the active site of an enzyme or other target. Unlike the normal substrate of the target enzyme, once bound they remain locked in position for an extended duration, preventing that enzyme molecule from translating it's substrate into some other product, and hence halting the pathway. An important note is that enzymes are often extremely efficient, and hence exist in very small concentrations. Because of this, overwhelming doses of such inhibitors are not required. The small size of such inhibitors also bestows certain thermodynamic advantages, as they require less energy to move around the body, degrade, etc.

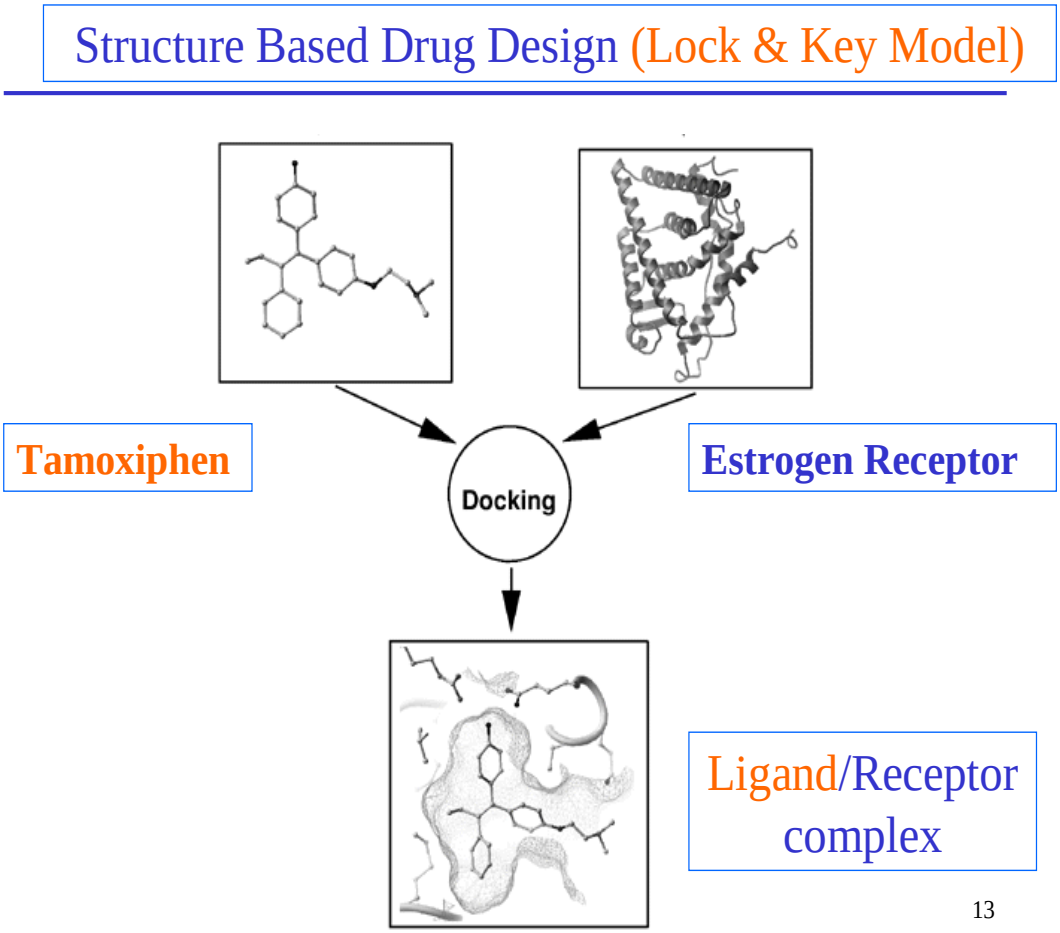
Overview of SMI selection and Fold Prediction

The small size of such inhibitors simplifies prediction of their 3D conformation, which makes them excellent candidates for structure based drug design (SBDD) initiatives. Traditional selection of SMIs has often been a somewhat random search, and structure based approaches attempt to look for patterns, or moieties, in an enzyme's 3D conformation which might strongly bind a certain structural feature of a drug. Characteristic clefts, hydrophilic and/or phobic pockets, and even protruding hydrogens or oxgens can be identified as one half of a given functionality. From that knowledge, a scientist can work with automation tools to try and find features such as sheets, helices, fingers, hydrophobic protrusions and hydrogen bonding partners to fill the target spaces. Given the enormity of even a trivial genome, SBDD is required to work on a well chosen small subset of candidates. The 'easy' half of the problem is the examination of the enzyme, as we know which sequence we wish to evaluate. When searching for a key, large genome databases are often screened with the various search engines earlier seen for sequences similar to a library of known local conformations. Many other techniques are under development for determining and/or examining the three dimensional conformation of a protein, and many show significant promise for small molecules such as SMIs. One feature of protein folding which has given drug hunters a source of relief is the existence of 'fold families'. Primary sequences with as little as 25-30% homology can be designated 'remote homologues' (although the prediction of folding from this similarity is unreliable, it is the basis for most comparative analysis). When looked at in terms of overall functional shape, proteins appear to fall into a small number of fold families. The exact number of families is contested in the literature, and is growing, albeit much slower than the number of proteins recorded is growing. A scientist attempting to find a protein which folds into a particular shape need only use BLAST, FASTA, or some other search engine to look for sequences in a database which are likely to take on this general conformation. Optionally, if he/she has ascertained which

small-scale motifs need to be found in the target drug, a search for natural products within a genome with the desired subunits and overall conformation can be done. These natural products can often then be altered slightly to delete their active components and increase their binding affinity for the target enzyme.

Protein-Ligand Docking

A key portion of drug design comes from the combinatorial chemistry approach. Drug companies are primarily concerned with what ligands will elicit a desired response in an individual. Often, this is accomplished with little or no understanding of the underlying chemistry. This can almost be viewed analogous to the classical image of a thousand monkeys hammering away at a thousand typewriters. Throw enough assays at the problem, and you are bound to find something that looks like a cure. But the key is that at the point where you've identified chemicals that produce the desired effect, you are no closer to understanding the cure. Often the exact mechanism will not be understood until well after the clinical phase. However, the drug development picture is one that trades off tolerance with response. Many of the high profile drugs, such as the ones used in treating HIV, have nasty side effects that are only barely offset by the benefits of taking them.



Often, the side effects are due to the lack of specificity the drug has for its intended target. If a chemical can be found that has a tighter affinity for the binding site of the original drug, then the new candidate can be given at lower (and hopefully less toxic) doses to yield the desired effect. Ideally, those proteins that are not intended targets for the drug will be less affected by the new drug. Taking this to the extreme, the perfect drug will interact with only the target protein and have a perfect affinity for the binding site only. This is the holy grail of any perfect drug design.

The basic idea is pretty simple. Once you figure out how an existing ligand docks with a protein, try to figure out what other molecules will dock in a similar fashion, hopefully even tighter than before. Taking this idea one step further, given a protein that you want to affect, find a molecule to activate or inhibit the active site. Ligand candidates are evaluated by some type of measure that attempts to account for the energy of the protein-ligand interaction that usually boils down to complex energy calculations of the system. Once you find the ligands that bind with the lowest energy, you have a refined set of drug candidates to test. There are

many complications that make this task very difficult however. First, a given a ligand and a protein, there are infinitely many orientations the two can take, any one of which is the ideal conformation. *How can the search be narrowed down to only those relevant conformations*? Even if a ligand fits perfectly in the active site, there is no guarantee that the protein has the required flexibility to allow the ligand to achieve that conformation. *How is the docking process modeled with respect to time*? Perhaps even more important is the calculation of the energy of the system.

With the advancement of genomics studies and bioinformatics tools, the view of drug design is going to change. Future drug development is going to start involving more direct approaches to solving the problem, allowing for the discovery of near-perfect drugs. Structure based drug design with involvement of genomic sequences of pathogen is showing more promise as a viable technique in the development cycle. Most of pharmaceutical industries are going to look at one facet of this type of drug design as it relates to get winning edge over other competitor industries. Drug discovery is no more a random process. There is a shift from descriptive discovery to predictive discovery. An important step in drug discovery is the cost-effective identification of lead molecules. Computer-assisted drug design (CADD), also called computer-assisted molecular design (CAMD), represents more recent applications of computers as tools in the drug design process. Computers are not substitute for a clear understanding of the system being studied, but an additional tool to gain better insight into the chemistry and biology of the problem. The technique is employed not only to predict the biological activity but also physicochemical and pharmaceutical properties prior to synthesis. Many of the large pharmaceutical companies have established internal bioinformatics groups whose purpose is to beat the competition to solutions of a problem that may give their company that crucial edge in producing the next major drug.

References:

ALIGN 5-1-1-DUPUWE/France http://www2.igh.cnrs.fr/bin/align-guess.cgi or http://genome.eerie.fr/fasta/align-query.html Applies the BLOSUM50 matrix to deduce the optimal alignment between two sequences.

Analysis and Annotation Tool for Finding Genes in Genomic Sequences 5-3-2-DcDPDFWE/USA http://genome.cs.mtu.edu/aat.html Identifies genes in a DNA sequence by comparing it to cDNA and protein sequence databases (including those at HGI, TIGR, dbEST, Swiss-Prot and nr).

BankIt (GenBank) http://www.ncbi.nlm.nih.gov/BankIt/

Beauty http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html Beauty is an enhanced BLAST search, returning output which predicts the function of the protein being tested.

Berkeley Fly Database 5-5-2-DDW (UI)/USA http://www.fruitfly.org/bfd/ Search for sequences by name or map location, and optionally view a clickable image of sequenced contigs aligned alongside the fly chromosomes. Searches can be limited to available sequences only. Retrieve P1, BAC or cosmid genomic clones, P element insertion lines, YAC, STS and more.

Berkeley *Drosophila* Genome Project BLAST Searches 5-4-3-DDFWE/USA http://www.fruitfly.org/blast/ Search for your sequence using the WU-BLAST 2.0

BLAST 2 Similarity Search (EMBNet) 5-3-2-PUWE/Switzerland http://www.ch.embnet.org/software/frameBLAST.html WU-BLAST 2.0 similarity searches.

BLAST2 Search with Post-processing (EMBL) 4-3-2-PUW/Germany http://dove.embl-heidelberg.de/Blast2/ WU-BLAST 2.0 search with post-processing. algorithm for *D. melanogaster* sequence data, including EST's, genomic sequences, STS's or sequences derived from them, P element insertion sites and transposons.

Blitz http://www.ebi.ac.uk/searches/blitz.html

Colour INteractive Editor for Multiple Alignments (CINEMA) http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.1/ A comprehensive and popular site. Allows the user to visualise and manipulate aligned protein sequences. *Makes use of Java. I recommend you access this site from a fast workstation!*

Compute the Theoretical pI and Mr 5-1-2-PUW/Switzerland http://www.expasy.ch/ch2d/pi_tool.html Calculates the theoretical pI or molecular mass of a protein, whose sequence is entered by the user, or referred to as a SWISS-PROT or TrEMBL entry.

DNA Databank of Japan (DDBJ) http://www.ddbj.nig.ac.jp

EBI Primers Database http://www.ebi.ac.uk/primers_home.html

Entrez http://www.ncbi.nlm.nih.gov/Entrez/ Start here for pretty much anything

European *Drosophila* Genome Project BLAST server 5-4-3-DDPDFWE/UK http://edgp.ebi.ac.uk/www-blast.html Search using the WU-BLAST 2.0 (gapped aligned) or the original BLAST (which does not allow gaps). The database includes *Drosophila* genomic data, EST's, STS's, P element sites, transposons, repeats, and proteins.

European Molecular Biology Laboratory (EMBL) http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html Cambridge, UK.

FASTA 3 (EMBL) http://www2.ebi.ac.uk/fasta3/ FASTA 3 similarity search.

FASTA http://www2.igh.cnrs.fr/bin/fasta-guess.cgi FASTA similarity search and a clean, basic and simple interface.

Forward and Reverse Translation 5-5-2-DUPUFWE/UK http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml Translate a protein sequence into a genomic DNA sequence, and vice versa. *This is a WWW interface to the pgwise software application. Those who are proficient with this package may like to take advantage of it's extra capabilities by adding execution critera.*

GenBank http://www.ncbi.nlm.nih.gov/Web/Search/index.html GenBank at the National Center for Biotechnology (NCBI) of The National Library of Medicine (NLM) at The National Institutes for Health (NIH) campus, USA.

Genome Sequence DataBase (GSDB) http://seqsim.ncgr.org/ The National Center for Genome Resources, Genome Sequence Database. The server is a supercomputer with genomic algorithm accelleration. These are often specialised databases.

Genome Database (GDB)http://www.hgmp.mrc.ac.uk/gdb or http://gdbwww.gdb.org/ *Funding for this project has been withdrawn. This valuable database will remain online, but it should be noted that no new entries will be recorded after 31st July 1998.*

GENATLAS (*) **5-5-2-DUWI/France** http://bisance.citi2.fr/GENATLAS/ A comprehensive, easy to use site. Search gene, marker or phenotype or linkage databases. Useful, relevant links provided in the results. The user can also locate the desired gene from a graphical clickable map of disease related or other mapped genes on a chromosome.

Gene Cards http://bioinfo.weizmann.ac.il/cards/ A very useful site providing comprehensive information and links. Direct links to GenBank, SWISS-PROT and MedLine. Includes synonyms, similar genes in other organisms, gene products and details about disorders.

HGMP-RC Primers Database http://www.hgmp.mrc.ac.uk/local-data/Primers.html

Human Genome Map Database (HuGeMap) http://www.infobiogen.fr/services/Hugemap Genetic and physical maps of the human genome. Connected with the gene radiation hybrid mapping database RHdb.

Human Gene Mutation Database 4-2-2-DUW/UK http://www.uwcm.ac.uk/search/mg/allgenes? Enter a GDB accession number, disease name, gene name or symbol to retrieve well presented information about the different mutation types, information and links.

Molecular Probe Data Base (MPDB or MOLPROBE) http://www.biotech.ist.unige.it/interlab/mpdb.html

Multiple Sequence Alignemnt with MAP 4-3-2-DcUPUFWE/USA http://genome.cs.mtu.edu/map/map.html Calculates the global alignment of DNA or protein sequences using an algorithm which computes the best overlapping alignment without penalising terminal gaps. Long internal gaps in short sequences are not penalised.

Multiple Sequence Alignment with Hierarchial Clustering 5-5-3-PUW/France http://www.toulouse.inra.fr/multalin.html Sequence alignment with a colour output where differing or similar amino acids in the alignment can be highlighted.

MITOMAP http://infinity.gen.emory.edu/mitomap.html Mitochondrial DNA database.

Network Protein Sequence Analysis 5-4-1-PUW/France http://pbil.ibcp.fr/NPSA/npsa_clustalw.html ClustalW multiple sequence alignment.

Nucleotide to Protein (ExPASY) 5-2-1-DUW/Switzerland http://www.expasy.ch/tools/dna.html Translates a nucleotide sequence (DNA/RNA) into a protein sequence (amino acids).

Nucleotide to Protein (EMBL) 5-4-1-DUW/UK http://www.ebi.ac.uk/contrib/tommaso/translate.html Translates a nucleotide sequence into a protein sequence.

ORF Finder http://www.ncbi.nlm.nih.gov/gorf/gorf.html Finds likely open reading frames in a sequence.

Pairwise Sequence Alignment 3-4-2-DcUPUFWE/USA http://genome.cs.mtu.edu/align/align.html Computes the global alignment between two sequences. Compare DNA with DNA, cDNA or protein. For DNA and cDNA, settings (gap open penalty, gap extension etc.) can be defined.

Protein Mutation Database *Check this site!* http://www.genome.ad.jp/dbget-bin/www_bfind?pmd COMMENTS

Protein and cDNA Translation 5-4-2-DcUPU/UK http://www.sanger.ac.uk/Software/Wise2/protein2cdna.shtml

PIR http://www_nbrf.georgetown.edu/pir/ Four databases: PIR1 is the most comprehensive with entries classified and annotated. PIR4 is the least comprehensive, with un encoded or untranslated entries.

Pratt Search (EMBL) 4-4-2-PU http://www2.ebi.ac.uk/pratt/ Interactively identifies conserved patterns from a series of user-entered unaligned protein sequences.

PROSITE (via EBI) http://www2.ebi.ac.uk/ppsearch/ Pattern search to identify conserved functional amino acid motifs. Scans a sequence against PROSITE (the primary motif database) with a graphical output.

PROSITE http://expasy.hcuge.ch/sprot/prosite.html The primary motif database. Take care! Motifs are often short, and a large number of false positives should be expected! Options are available to exclude those which most commonly lead to false results.

Protein Motif Fingerprints http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS.html

Protein Databank http://www2.ebi.ac.uk/pdb/index.shtml y (viewer required) or viewed in a hypertext browser window (e.g. Netcape). The structures are experimentally determined by X-ray crystallography and nuclear magnetic resonance (NMR) imaging.

Protein Colourer 5-2-1-PUW/UK http://www.ebi.ac.uk/htbin/visprot.pl Colour a protein sequence (raw text or SWISS-PROT Acc. No.) by properties e.g. hydrophobicity.

Primer 3 5-5-2-DUW/Norway http://www2.no.embnet.uo.org/primer/primer3.cgi? Select PCR primers for your nucleotide sequence.

Random Protein Sequence Generator 5-5-2-PUW/Switzerland http://www.expasy.ch/sprot/randseq.html Random protein sequence generator! Output in FASTA format (the format most commonly required by bioinformatics search sites).

Residue Periodicity Review this site http://o2.dbuoa.gr/FT/ Study the periodicity of residues in a protein sequence.

Sequence Retrieval System (SRS) http://srs.hgmp.mrc.ac.uk/

SWISS-PROT http://expasy.hcuge.ch/sprot/sprot-top.html A database of protein sequences, translated from the EMBL genomic database. Protein sequences have been checked and annotated.

Translated EMBL (TrEMBL) http://www.expasy.ch/sprot/sprot-top.html Database of all the protein coding regions stored in the EMBL database. Comprehensive, but (generally) at the cost of poor annotation.

The Institute of Genomic Research Databases http://www.tigr.org/tdb/tdb.html Many databases including microbial, parasites, human, human cDNA, mouse, rat, *Arabidopsis*, zebrafish, and others.

The International Immunogenetics Database (IMGT) http://imgt.cnusc.fr:8104 Contains expertly annotated sequences and alignment tables for Ig, TCR and MHC sequences.

TIGR HGI Gene Expression Data 4-2-1-DDcW/USA http://www.tigr.org/tdb/hgi/searching/hgi_xpress_search.html Search for tissue specific transcripts e.g. "lung". cDNA libraries can also be searched.

The Sanger Centre Database Search Services 5-2-5-DDPDFWE/UK -Clean, simple design. http://www.sanger.ac.uk/DataSearch/ BLAST and WU-BLAST 2.0 searches which can be refined to finished and/or unfinished genomic sequences.

UTR Home Page http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/ Internet resources for sequence analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. Includes specialised UTR databases and tools for analyses of UTR regions.

VSNS BioComputing Division Multiple Alignment Resource Page http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/ An excellent, comprehensive resource for multiple sequence alignment, software and tutorials.

Via OMIM You can search for a 'disease gene' at OMIM. Click the "DNA" button in the results display and follow the link to the mRNA sequence. Note the absence of U (uracil): this sequence is referred to in GenBank reports as mRNA, but the sequence is a cDNA sequence. http://www.nih.gov

Web Cutter - Restriction Enzyme Mapping Utility http://rna.lundberg.gu.se/cutter2/ Map restriction enzyme sites on your sequence.