rithms. Since the templates fail to provide the structural information of many protein side chains, these can therefore not be assimilated in a first attempt and must be added later by further simulation. The amount of side chains required to be constructed depends on the degree of sequence identity between target and template protein. The method has its major weakness regarding protein divergences, which are not connected to any typical homologous family. [137]

III. Data and Bioinformatics for proteomics

1. Databases

To catalogue all human proteins and reveal their function and interaction is an immense challenge for scientists. The number of databases providing biological information is steadily increasing.[138] A new discipline, bioinformatics, has emerged to utilize the information in these databases for a better understanding of biological processes. Bioinformatics scientists work to interpret experimental data. Several sequence and sequence-related databases are available for public use. Additionally, a number of specialized databases exist which focus on a single enzyme, protein family or disease.[139] For amino acid sequences, the databases differ in their content. SWISS-Prot, established in 1986, strives to provide a high level of protein annotation with several cross-links to other databases.[140] Since 2003, it has been carried out by the UniProt Consortium, a collaboration between the Swiss Institute of Bioinformatics (SIB) and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown

137 Maggio, Edward T./Ramnarayan, Kal, Recent Developments in Computational Proteomics, 19 Trends in Biotechnology 2001, 266, 271.

138 Detailed overview of bioinformatics techniques of importance in protein analysis is provided by: Persson, Bengt, Bioinformatics in protein analysis, In: Proteomics in Functional Genomics - Protein Structure Analysis; Jollès, P./Jörnvall, H. Ed. Basel, Boston, Berlin, 2000; 215. Access to most protein databases is free. Recently, however, many providers stopped granting open access and started requiring licenses from commercial users. In the future, even academic users might have to register and pay. Goodman, Phillip, Access Ability, Genome Technology 2004, 21. Carugo, Oliviero/Pongor, Sàndor, The Evolution of Structural Databases, 20 Trends in Biotechnology 2002, 498 emphasize that the evolution of structural databases has been driven by the practical application of structural knowledge.

139 Links to biologically relevant databases are available at the web pages of EBI (European Bioinformatics Institute, Hinxton, England; http://www.ebi.ac.uk), the University of Geneva, Switzerland (http://www.expasy.ch), and NCBI (National Center for Biological Information, Bethesda, MD, USA; http://ncbi.nlm.nih.gov).

140 An annotation gives a narrative description to the formal structure of a protein, Carugo, Oliviero/Pongor, Sàndor, The Evolution of Structural Databases, 20 Trends in Biotechnology 2002, 498, 498.

University Medical Center's Protein Information Resource (PIR).[141] TrEMBL, which stands for Translated EMBL, is another large protein database, carried out by EBI/EMBL[142]. It is constituted in the same format as Swiss-Prot. It consists of computer translations of genetic information contained in the EMBL Nucleotide Sequence Database,[143] which are not yet integrated in SWISS-PROT. PIR (Protein Information Resource),[144]a U.S. protein related organization, has established the Protein Sequence Database (PSD) that contains functionally annotated protein sequences, which grew out of the "Atlas of Protein Sequence and Structure" (1965-1978) edited by Margaret Dayhoff. [145]Apart from that, GenPept is a database, which contains translated protein-coding sequences, which were produced by translating open reading frames from GenBank, the NIH genetic sequence database.[146] Furthermore, 3-D structures of biological macromolecules are collected in the Protein Data Bank (PDB) maintained by the Research Collaboratory for Structural Bioinformatics (RCSB).[147] More recently, another data resource has been initiated in In-

141 Available at http://www.expasy.org/sprot/sprot_details.html, last checked on January 21, 2008.

142 The European Molecular Biology Laboratory is a non-profit organization and a basic research institute funded by public research monies from 20 member states. Research at EMBL is carried out by approximately 80 independent groups covering the whole spectrum of molecular biology. The Laboratory is divided into five units: the main Laboratory in Heidelberg, and Outstations in Hinxton (the European Bioinformatics Institute), Grenoble, Hamburg, and Monterotondo in the Rome region. The key issues of EMBL's work are: to perform basic research in molecular biology, to train scientists, students and visitors at all levels, to provide crucial services to scientists in the member states, and to develop new instruments and methods in the life sciences, and technology transfer. The European Bioinformatics Institute (EBI) is associated with EMBL.
See http://www.embl-heidelberg.de/aboutus/index.html, last checked on January 20, 2008.

143 The EMBL Nucleotide Sequence Database (also referred to as EMBL-Bank) is Europe's primary nucleotide sequence collection. The DNA and RNA sequences are mainly obtained from submissions of individual researchers, genome sequencing projects, and patent applications. The database is maintained in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ), available at http://www.ebi.ac.uk/embl/, last checked on January 21, 2008.

144 Further information on the Protein Information Resource center is available at: http://pir.georgetown.edu/, last checked on January 21, 2008.

145 Dayhoff, M. O., Eck, R. V. and Park, C. M. Atlas of Protein Sequence and Structure, Vol. 5, 75, London 1979, published by National Biomedical Research Foundation (NBRF).

146 GenBank contains a collection of all publicly available DNA sequences, which are released at the NCBI ftp site, available at ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt, last checked on January 21, 2008. Open Reading Frames (ORFs) are DNA protein-coding sequences, which devoid of stop codons and are therefore suitable for RNA polymerase, see Alberts, Bruce/Johnson, Alexander.Lewis, Julian, Molecular Biology of the Cell (4th ed.), New York 2002, 110-111.

147 Available at http://www.pdb.bnl.gov.; the PDB is common ancestor of all structural databases. It was established in 1971. Over the years, the quantity, phenotype and quality of the deposited structures have changed due to new experimental techniques. A good description is provided by Carugo, Oliviero/Pongor, Sàndor, The Evolution of Structural Databases, 20 Trends in Biotechnology 2002, 498, 499.

dia, where the "Institute of Bioinformatics" (founded in 2002) works to establish the "Human Protein Reference Database" (HPRD)[148] – "a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks, and disease association for each protein in the human proteome." Most of the protein annotation data it contains is more or less redundant with SWISS-Prot; the interaction data, however, goes far beyond. It is set apart by manual curation, which means a reliable way to control quality compared to other databases that are created by automatic processes. So far, none of the existing databases can be considered an established standard, in fact all are still in their early stages. Hence, the existing variety offers scientists the possibility of choosing instead of imposing one database by default. The latest accomplishment of a 'Human Proteome Atlas for Normal and Disease Tissue, established by the Swedish Human Proteome Resource (HPR) program, funded by the Knut and Alice Wallenberg Foundation, however, represents a highlight of proteomics endeavor that might set new standards for proteomic research.[149]

2. Cross-linking of database information

In 2000, the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI) founded the Human Proteomic Initiative (HPI) with the major goal to annotate, describe and distribute a large amount of information concerning human protein sequences to the life science community. Being aware of the huge complexity of the proteome, the initiative aims to annotate all known human sequences according to the quality standards of SWISS-Prot. These standards include more than 9000 annotated human sequences associated with about 23200 literature references; 22600 experimental or predicted PTM's, 2800 splice variants and 15100 polymorphisms. The interpretation of all known human sequences for each known protein includes a wealth of information. It refers to the description of its function, domain structure, subcellular location, post-translational modifications, and variants or similarities to other proteins. The HPI project contains a number of subcomponents, such as

- Analysis of all known human proteins,
- analysis of mammalian orthologs[150] of human proteins,
- analysis of all known human polymorphisms at the amino acid level,

148 Genome Technology 10, 2003, 16.

149 The protein atlas aims to demonstrate the expression and localization of proteins in large variety of normal tissue and cancer cells. The basic concept of the resource centre is to produce antibodies to human target proteins. The antibodies are subsequently used for functional analysis of the corresponding proteins in numerous further platforms. See http://proteinatlas.org/, last checked on January 21, 2008.

150 Proteins orthologs are proteins that have evolved from the same inherited region.

- analysis of all known post-translational modifications in human proteins,
- tight links to structural information, and the
- classification of all known vertebrate proteins.[151]

3. Database screening and drug design

The structure-based screening approach aims to identify a subset of an existing or a virtual library of compounds with an enhanced probability of binding. Typically, the first step of drug development is the selection of an adequate target. When a high-quality structure is available for investigation, it can be used to screen databases or libraries of existing chemical substances. The screening process will then select those substances containing an array of chemical groups compatible with binding to the targeted binding pocket of the protein.

Compounds possessing the desired binding characteristics are analyzed in further tests. If the tested effect is confirmed, it is examined to determine whether the compound is pharmaceutically acceptable and toxicologically safe. If the drug succeeds, auxiliary substances are added for the final stage of the pharmaceutical process.[152] Drug development research using 3-D structure information can be illustrated using the graphical description depicted in figure 8.[153]

151 The web-page of HPI is available at http://www.expasy.org/sprot/hpi/hpi_desc.html, last checked on January 21, 2008.

152 Maggio, Edward T./Ramnarayan, Kal, Recent Developments in Computational Proteomics, 19 Trends in Biotechnology 2001, 266, 271. The process of drug discovery involves immense labors. For views from inside the pharmaceutical industry, see Mervis, Jeffrey, The Hunt for a New Drug: Five Views from the Inside, 309 Science 2005, 722, 722.

153 The figure was adopted from Masuoka, Kunihisa, Study on the Ways of Protection of Post-Genome Research Products, IIP Bulletin 2002, 84, 87.

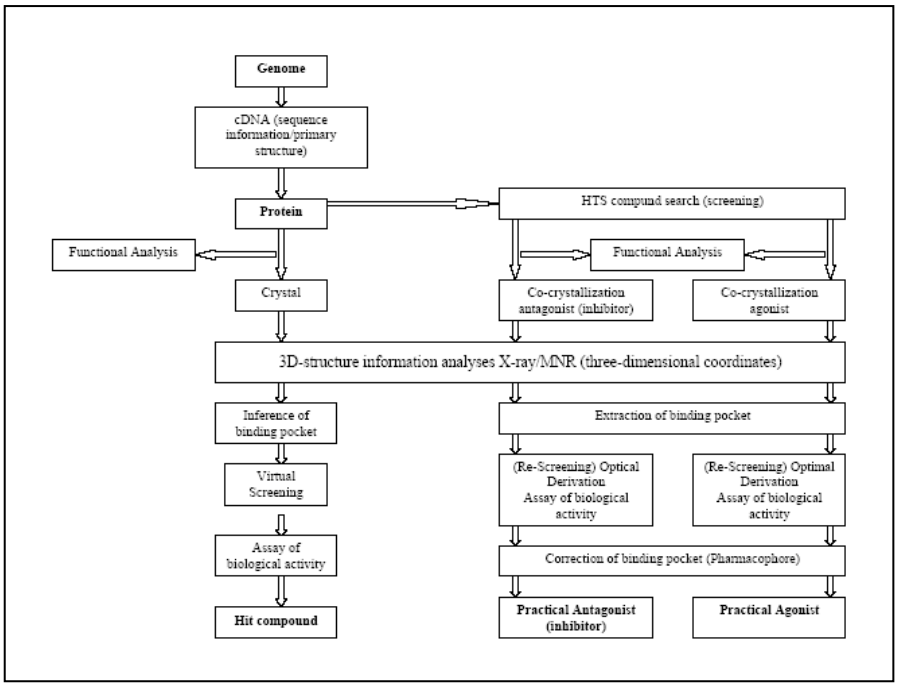

Figure 8: Stylized illustration of proteomic drug design

4. In-silico screening of binding pockets

In-silico[154] screening methods are methods that aim to scan chemical/ pharmaceutical compounds for new drug design. They involve the computerized simulation of the three-dimensional structure of a polypeptide. The simulated protein is then used to screen several pharmaceutical compound-related databases. In order to determine the pharmaceutical/chemical response of binding pocket properties the screening methods comprise several steps. These include the application of 3-D molecular modeling algorithm to the atomic coordinates of a protein, the determination of the spatial coordinates of binding pockets, and the electronic screening of candidate compounds against the spatial coordinates of the protein. The major goal is to identify compounds that can bind to the computerized protein. More precisely, the molecular model simulates the positions of heteroatoms in the amino acids, which form the binding pockets of the protein. It also includes information about hydrogen

154 From literally: 'in the computer'.

bonds. The coordinate data of the computerized protein is then incorporated in the database such that the interatomic distances between the atoms of the simulated protein is retrieved. In a further step, the distances between the bonding of different candidate compounds and the atoms that bind in the computerized protein model are compared. Thereby, it is possible to identify those candidate compounds that would theoretically form the most stable complex with computerized 3-D model.[155] The obtained ligands can efficiently be used for the development of new drugs.[156]

In-silico screening methods replace the traditionally used *in vitro*[157] methods, which were generally based on a 'trial and error' approach. The Human Genome Project and the improvement of protein analysis techniques has lead to a dramatic increase in the information to be interpreted. The method of in vitro research thus became too expensive and time consuming. However, since an *in-silico* screening is only hypothetical and based on simulated structures, it always requires *in vitro* testing of useful identified compounds in order to verify that the underlying technical problem of finding an appropriate agent has indeed been solved. Consequently, a biological evaluation of the obtained compounds is necessary.[158]

155 A good overview is provided by Gnanakaran, S./Nymeyer, Hugh/Portman, John/ Sanbonmatsu, Kevin Y./Garcia, Angel E., Peptide Folding Simulations, 13 Current Opinion in Structural Biology 2003, 168.

156 See note: Innovatives in-silico Verfahren beschleunigt Wirkstoffsuche, Transkript 2004, 28, emphasizing that most recent in-silico methods rely upon protein structure homology for the search of new compounds.

157 In an artificial environment outside a living organism or body, for example, testing conducted in the laboratory.

158 Lonati, Milena, Patentability of Receptors and Screening Methods: Does in silico Screening Pose New Legal Problems? Bioscience Law Report 2000/2001, 144, 145.