Biomolecular Nanothecnology: computer based approach

Carmelina Ruggiero
DIST University of Genoa

Molecular Systems knowledge:

Genomics

Proteomics

Molecular Systems

Improved experimental methods

- crystallography
- spectroscopy

etc

Computer methods

- computer modelling for proteomics
- data manipulation and analysis for proteomics and genomics

The study of the expression, location, interaction, function and structure of all the genes or proteins

Genomics

Proteomics

Relevant topics

- Genomics
- Proteomics
- Protein structure
- Case study

Genomics

Main challenges:

- predict and annotate the function of genes
- derive suitable abstractions to compare genomes at a higher then molecular level

Main goals : function prediction

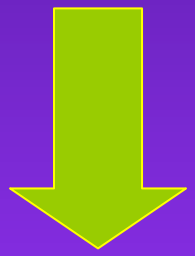
integration of methods that rapidly and efficiently identify similar sequences from databases, improving function prediction by homology

integrated database and software tools combined with expert knowledge

Genome Project

- There has been a revolution in molecular biology.
- Over the past 15 years, the entire genomes of several organisms, human included, have been sequenced.
- The task is now to make sense of the vast databases of information and work out how the instructions for building an organism are implemented.

Great amount of genetic information produced by genome projects.



combining tools and techniques of mathematics, computer science and biology it is possible to understand the significance of biological information

Proteomics

Proteomics is the analysis of genomic complements of proteins.

Main challenge

Proteomics tries to define the function, quantities and structures of large complements of proteins.

Goal

Understand the function and mechanism of proteins

Protein structure and engineering

Advances in the knowledge of protein structure and in the design and manufacture of proteins.

Methods:

- X-ray diffraction
- High resolution NMR
- Spectroscopy
- Theoretical methods for the study of molecular mechanics and dynamics
- Computational methods - both traditional and parallel

Functional annotation of proteins (protein sequence databases)

Protein Sequence → Function

From new genomes:

Automatic assignment based on sequence similarity

gene name, protein name, function

Best annotated protein databases: SwissProt, PIR-1
Now part of UniProt – unified protein knowledgebase

Functional prediction: computational analysis

- Cluster analysis of protein families (family databases)
- Use of sophisticated database searches
- Detailed manual analysis of sequence similarities

Protein analysis 1-2

- Comparative analysis allows us to find subtle sequence similarities in proteins that would not have been noticed otherwise
- Prediction of the 3D fold and general biochemical function is much easier than prediction of the exact biological (or biochemical) function.

Protein analysis 2-2

- Reaction chemistry often remains conserved even when sequence diverges almost beyond recognition
- Sequence database searches that use exotic or highly divergent query sequences often reveal more subtle relationships than those using queries from humans or standard model organisms (E coli, yeast, worm, fly).
- Sequence analysis complements structural comparisons and can greatly benefit from them

Secondary structure Prediction (1)

Applications:

- for the improvement of interpretation of low-resolution experimental results (prediction of the structure of the proteosome).
- Classification of protein structures
- Definition of loops (active sites)
- Use in fold recognition methods
- Improvements of alignments
- Definition of domain boundaries

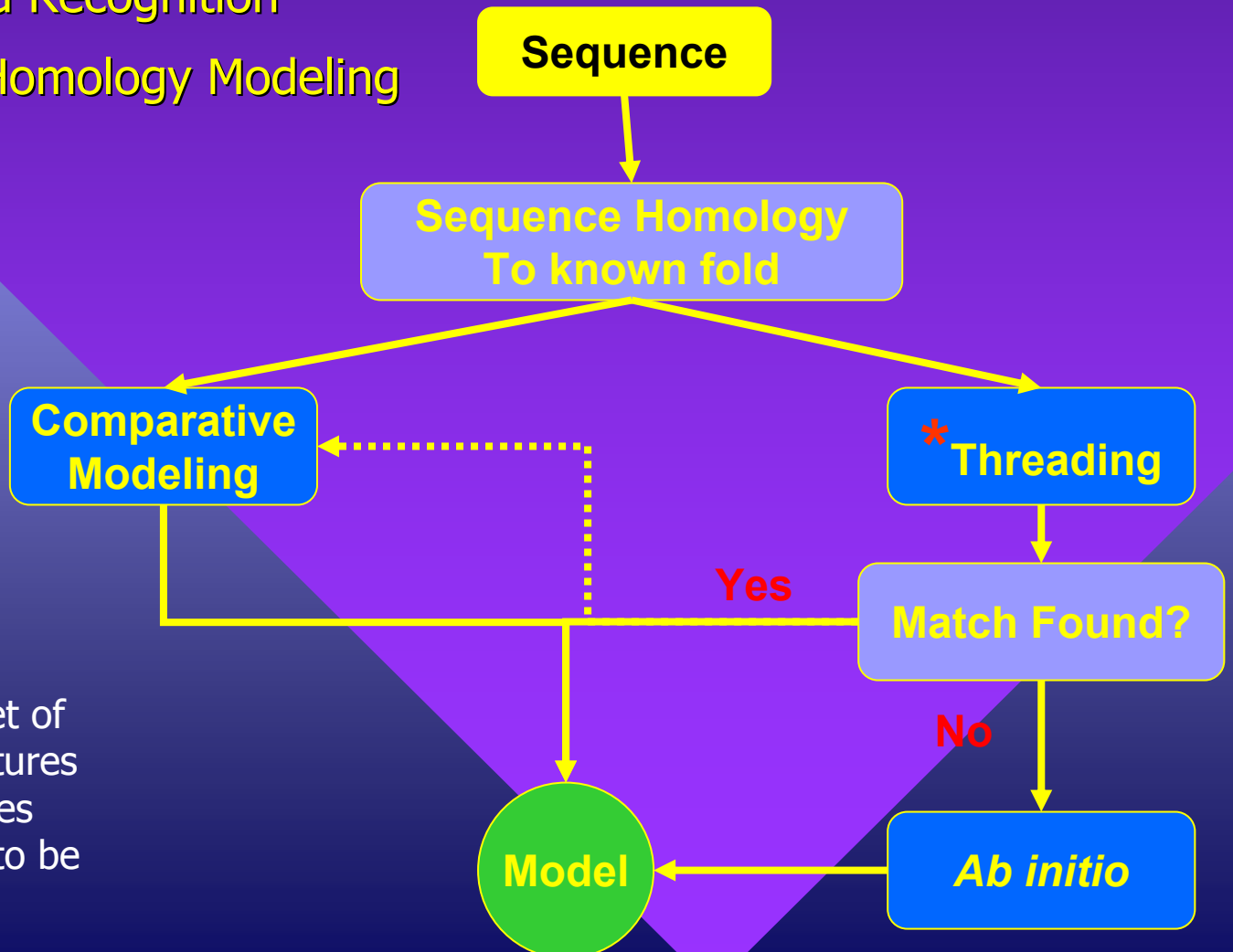
Secondary Structure Prediction (2)

- Simple alignments
  - Align to a close homolog for which the structure has been experimentally solved.
- Heuristic Methods (e.g., Chou-Fasman, 1974)
  - Apply scores for each amino acid an sum up over a window.
- Neural Networks
  - Raw Sequence (late 80's)
  - Position specific alignment profiles

Tertiary Structure Prediction (1)

Three major classes of algorithms

- Threading/Fold Recognition
- Comparative/Homology Modeling
- Ab initio



***Threading:**

looking through the set of currently known structures and identifying the ones which are most likely to be appropriate for the sequence of interest

Tertiary Structure Prediction (2)

Threading

Goals:

- Identify the *fold family* for a *target* sequence
- Compute the sequence-alignment between the target sequence and *template* structure
- Threading usually does not generate *all-atom models:*
  - Side chains atoms are usually not predicted
  - Unaligned residues are usually not predicted
  - Loop configurations may or may not be predicted

Tertiary Structure Prediction (3)

Threading

- Elements
  - Database of templates
  - Evaluation function
  - Search/Alignment algorithm
  - Confidence Score
- Threading models are generally not suitable for things like drug design
- Function prediction is only possible if the fold family is only associated with a single function

Tertiary Structure Prediction (4)

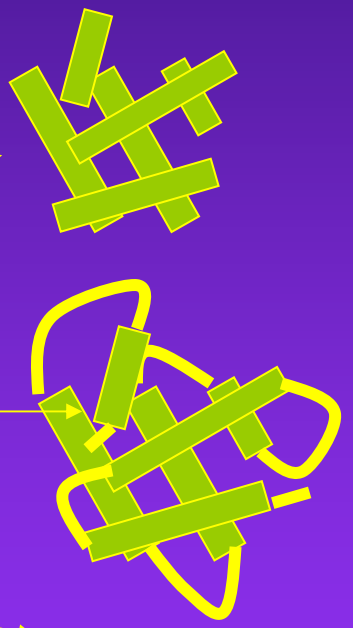
Comparative Modeling

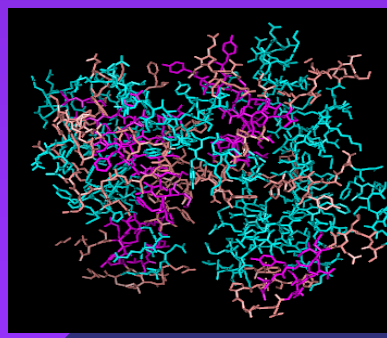
- Goals:
  - Construct an *all-atom model* of the target sequence from a template structure
  - Templates are chosen from a database of structures
    - Selection criteria
      - Sequence homology
      - Secondary structure prediction
      - Threading
- Elements
  - Database of templates
  - Evaluation function
  - Search/Alignment algorithm

Tertiary Structure Prediction (5)

Steps for **comparative models:**

1. Template Selection and alignment
   - Similar to threading
2. Partial backbone generation
3. Loop modeling
4. Side-chain modeling
5. Model Refinement
6. Model Evaluation





CM models can, in theory, be used for anything that an experimentally solved structure is used for

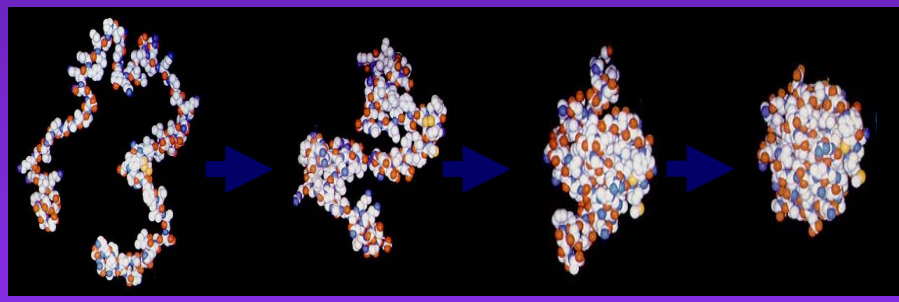
- Drug design
- Modeling protein-protein interactions

Tertiary Structure Prediction (6)

Ab initio

Goals:

- Construct an *all-atom model* of the target sequence *without* a template structure



Two subcategories

- Physical modeling
  - Simulated protein folding
- Statistical modeling
  - Predictions are made using a combination of statistical models relating sequence to structure
  - Statistical models are typical mined from a database of structures

Elements

- Evaluation function
- Search algorithm

Protein docking (1)

- Docking attempts to find the "best" matching between two molecules.
- Given two biological molecules
  - Do they interact ?
  - If so, what is the orientation that maximizes the interaction while minimizing the total "energy" of the complex.

Protein docking (2)

- Extreme relevance in **cellular biology**, wher function is accomplished by proteins interacting wit themselves and with other molecular components

- Key to rational **drug design**: docking

  find inhibitors for specific targets
  design new drugs.

Increasing importance as the number of proteins whose structure is known increases

Protein docking (3)

Types of Docking studies

- Protein-Protein Docking
  - Both molecules usually considered rigid
  - 6 degrees of freedom
  - First apply steric constraints to limit search space, then examine energetics of possible binding conformations
- Protein-Ligand Docking
  - Flexible ligand, rigid-receptor
  - Search space much larger
  - **Either** reduce flexible ligand to rigid fragments,

    or search the conformational space using monte-carlo methods or molecular dynamics

Protein docking (4)

Some techniques:

Surface representation: efficiently represents the docking surface and identifies the regions of interest (cavities and protrusions):

Surface matching: matches surfaces to optimize a binding score

Protein docking (5)

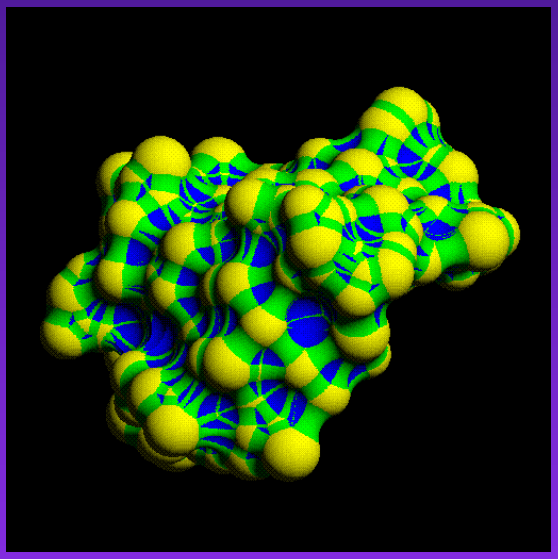
Surface representation

- *Connolly surface*

* Each atomic sphere is given the van der Waals

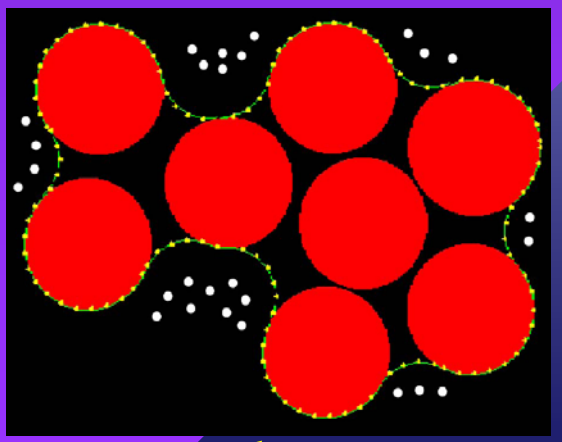
  radius of atom

* Rolling a Probe Sphere over the Van der Waals Surface leads to the Solvent Reentrant Surface or **Connolly surface**



- *Lenhoff technique*

* Computes a "complementary" surface for the receptor instead of the Connolly surface, i.e. computes possible positions for the atom centers of the ligand



Atom centers of the ligand

Atom centers of the ligand

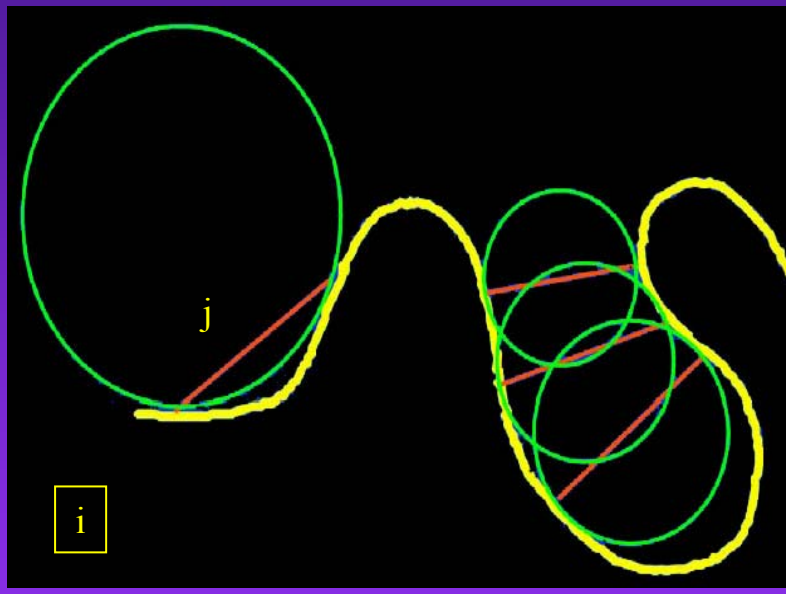
Protein docking (6)

Clustered-Spheres

Uses clustered-spheres to identify cavities on the receptor and protrusions on the ligand

Compute a sphere for every pair of surface points, i and j, with the sphere center on the normal from point i

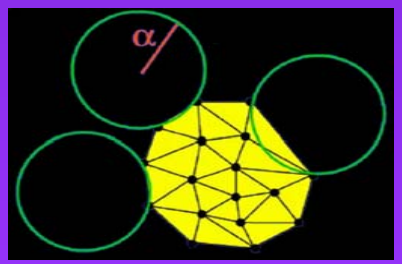
Regions where many spheres overlap are either cavities (on the receptor) or protrusions (on the ligand)

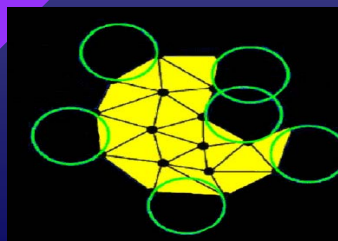


Alpha Shapes

* Formalizes the idea of "shape"

* In 2D an "edge" between two points is "alpha-exposed" if there exists a circle of radius alpha such that the two points lie on the surface of the circle and the circle contains no other points from the point





Why structural proteomics?

- To study proteins in their active conformation.
  - Study protein: drug design
- Proteins that show little or no similarity at the primary sequence level can have strikingly similar structures.

Drug discovery and biomolecular nanotechnology

The combination and integration of biomolecular and nanotechnology are crucial in drug discovery.

The use of complementary experimental and computer techniques increases the chances of success in many stage of the discovery process:

- identification of new targets,
- elucidation of their functions,
- development of compounds with desired properties.

Aim: to better understand

- cellular expression,
- family relation-ships,
- structure,
- function of proteins,
- to evaluate their potential as drug targets.

An application example:

genomics and proteomics for drug disegn

A view of biomolecular nanotechnology in drug discovery

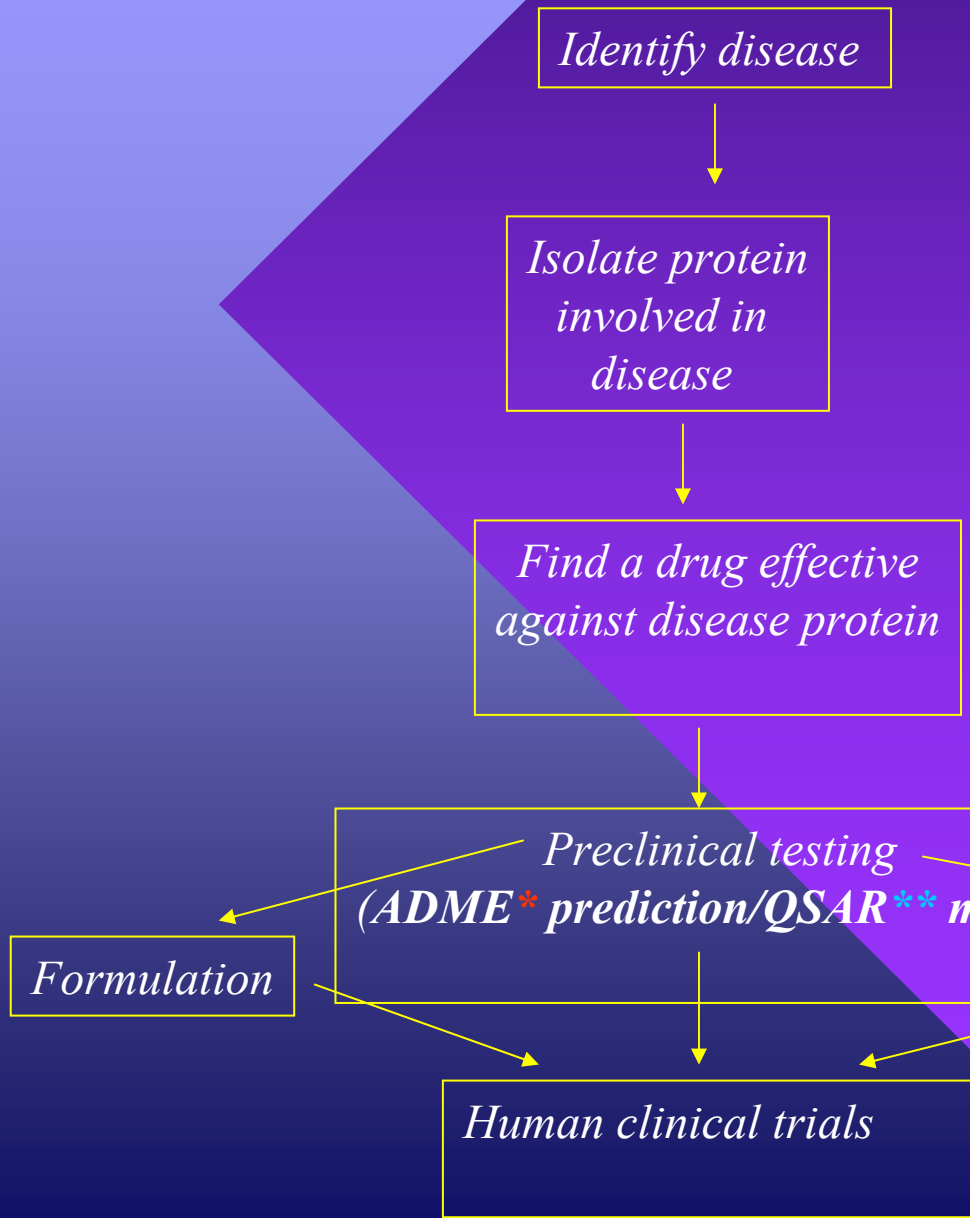
Several areas or stages of drug discovery are well complemented by biomolecular efforts.

Integration of biomolecular nanotechnology and chemo-informatics ⇒ supporting drug discovery programs at different levels:

- data management,
- database mining,
- novel design,
- discovery tools.

Consequences of this integration: "re-rationalization" of drug discovery research as predictive methods and introduction of frequently use of computational models.

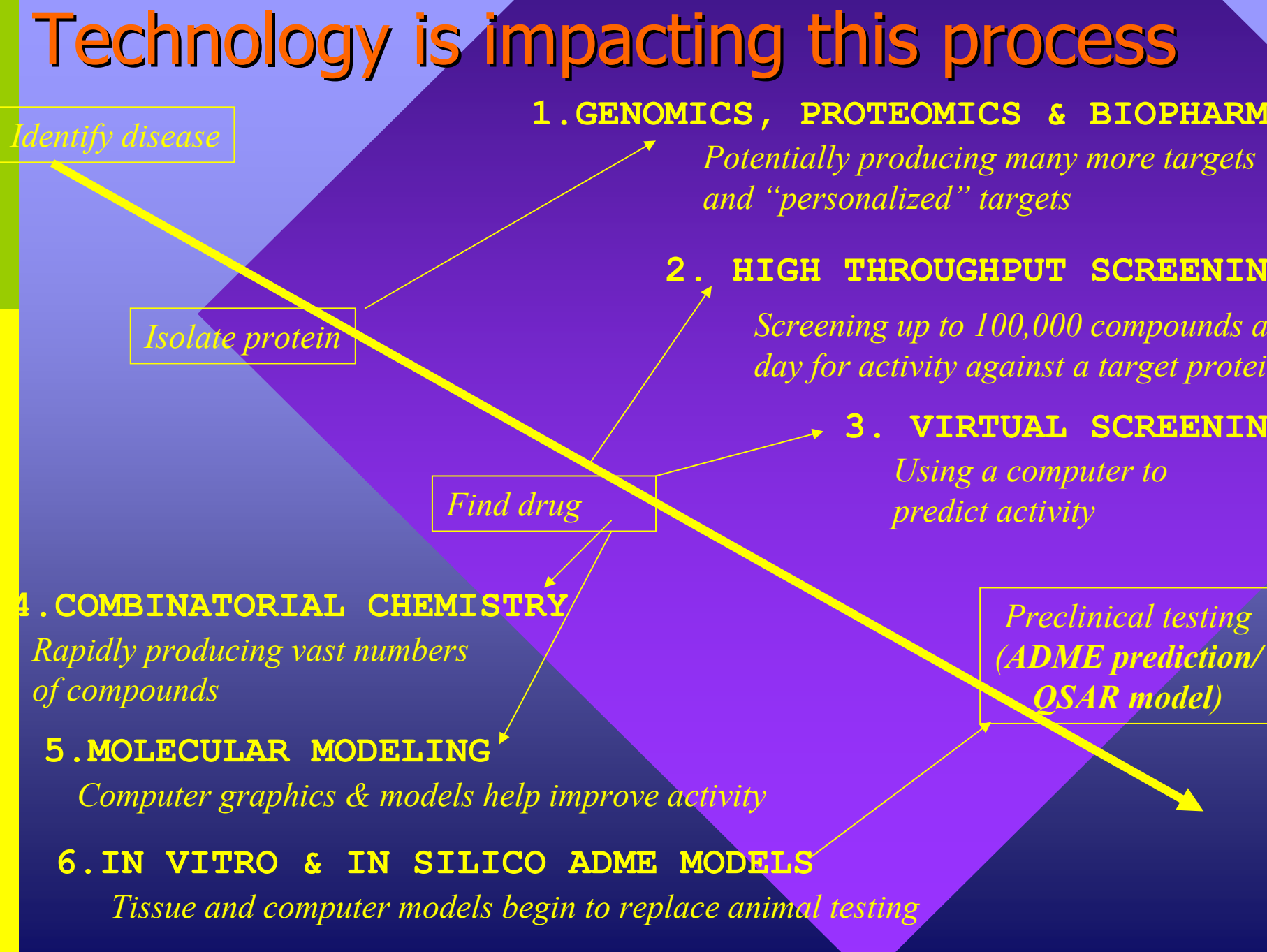
Drug Discovery & Development



***ADME**:* Absorption, Distributio
Metabolism, Excretion) technique
help model how the drug will likel
act in the body

***QSAR**:** Quantitative Structure
Activity Relationship):
mathematical models
approximating the often complex
relationships between chemical
properties and biological activities
of compounds

Technology is impacting this process



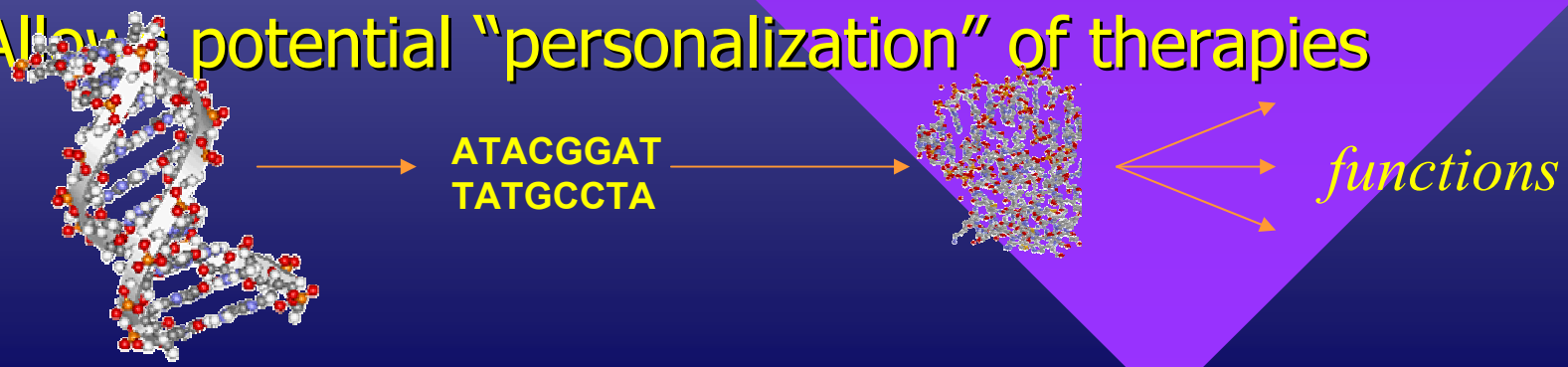
1. Genomics, Proteomics & Pharmacogenomics Discovery 1-3

Genomics is fast-forwarding understanding of how DNA, genes, proteins and protein function are related, in both normal and disease conditions

Human genome project has mapped the genes in human DNA

Hope is that this understanding will provide many more potential protein targets

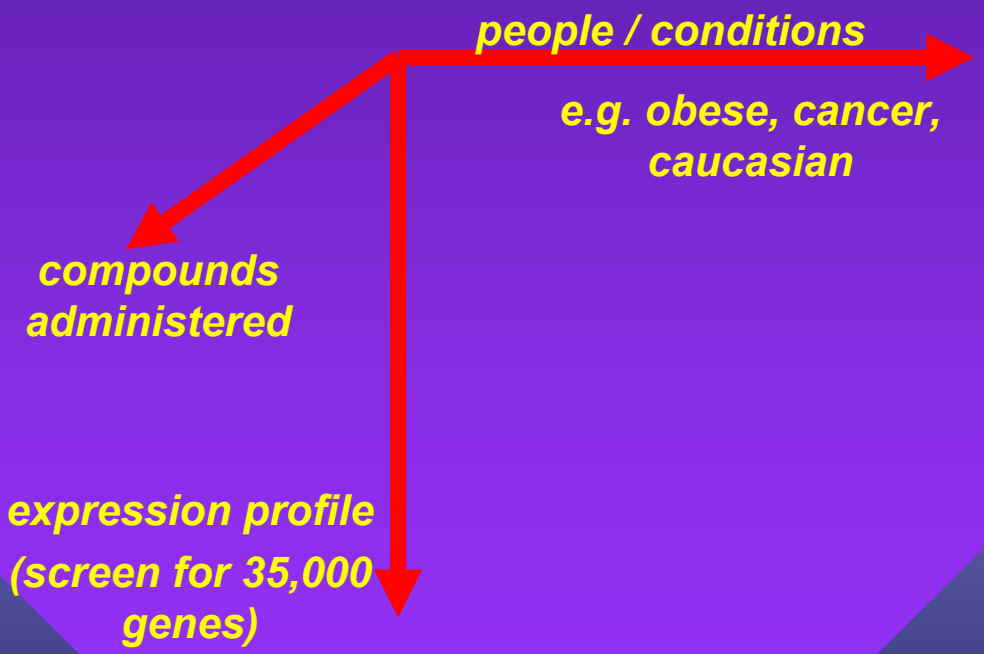
Allows potential "personalization" of therapies



L. Genomics, Proteomics & Pharmacogenomics Discovery 2-3

Gene chips":

allow to look for changes in protein expression for different people with a variety of conditions, and to see if the presence of drugs changes that expression



Makes possible the design of drugs to target different phenotypes

1. Genomics, Proteomics & Pharmacogenomics Discovery 3-3

Biopharmaceuticals:

Drugs based on proteins, peptides or natural products instead of small molecules (chemistry)

Pioneered by biotechnology companies

Biopharmaceuticals can be quicker to discover than traditional small-molecule therapies

Biotech now paring up with major pharmaceutical

2. High Throughout Screening 1-2

- Application:
  - Discovery
  - Hit development
  - Preliminary assessment of metabolism & toxicity
- Components:
  - Test substance supply
  - Bioassay development & implementation
  - Informatics

2. High-Throughput Screening 2-2

Drug companies have millions of samples of chemical compounds

High-throughput screening can test 100,000 compounds a day for activity against a protein target

May be tens of thousands of these compounds will show some activity for the proteins

The chemist needs to intelligently select the 2 - 3 classes of compounds that show the most promise for being drugs to follow-up

3. Virtual Screening

Build a computational model of activity for a particular target

Use model to score compounds from “virtual” or real libraries

Use scores to decide which to make, or pass through a real screen

4. Combinatorial Chemistry

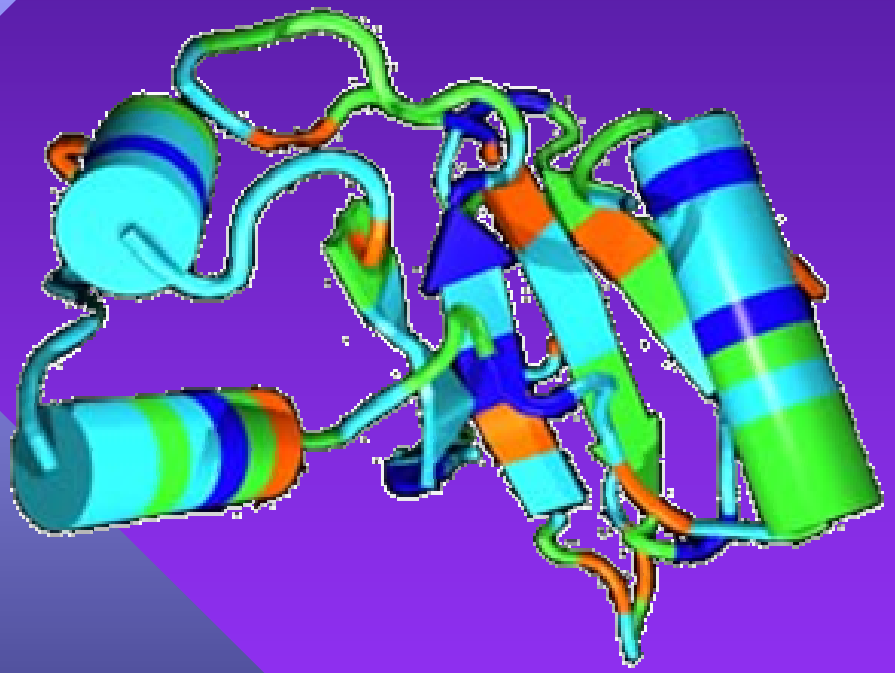
Large numbers of different molecules can be created by combining molecular "building blocks very quickly.

Involves a "scaffold" molecule, and sets of compounds which can be reacted with the scaffold to place different structures on "attachment points".

Issues:

1. Which R-groups to choose
2. Which libraries to make
   * "Fill out" existing compound collection?
   * Targeted to a particular protein?
   * As many compounds as possible?
3. Computational profiling of libraries can help
   * "Virtual libraries" can be assessed on computer

5. Molecular Modeling



- 3D Visualization of interactions between compounds and proteins
- "Docking" compounds into proteins computationally

6. In vitro & in silico ADME models

Traditionally, animals were used for pre-human testing.

Animal tests are expensive, time consuming and ethically undesirable

ADME (Absorption, Distribution, Metabolism, Excretion) techniques help model how the drug will likely act in the body

These methods can be experimental (*in vitro*) using cellular tissue, or *in silico*, using computational models.

In vitro models:

- Based around real tissue samples, which have similar properties to those in the body;
- Cuts down animal tests, by acting as a "pre-screen";

ADME Models

2. *In silico models:*

- Computational methods can predict compound properties important to ADME, e.g.
  - LogP, a liphophilicity measure
  - Solubility
  - Permeability
  - Cytochrome p450 metabolism

- Means estimates can be made for millions of compounds, helping reduce "attrition" – the failure rate of compounds in late stage.

The implementation of in silico tools for the ADME evaluation requires the prediction of physicochemical properties applying fast predictive **QSAR (Quantitative Structure-Activity Relationship) models**

What is QSAR?

QSAR's (Quantitative Structure Activity Relationship): mathematical models approximating the often complex relationships between chemical properties and biological activities of compounds.

Objectives: to allow prediction of biological activity of untested and unavailable compounds.

QSAR methods use databases of group contributions from which properties can be calculated for formulations made up of compounds that were in databases.

Building accurate QSAR prediction models for drug discovery tends to be very challenging.

The two largest challenges:

1. the relevance of the data;
2. the learning algorithm used to build the model.

QSAR and Drug Design

Compounds + biological activity

New compounds with improved biological activity

General Procedure of QSAR

- Select a set of molecules interacting with the same receptor with known activities.
- Calculate features (e.g. physicalchemical properties, etc., 2D, 3D)
- Divide the set to two subgroups: one for training and one for testing.
- Build a model: find the relations between the activities and properties (regression problem, statistic methods, machine learning approaches, etc).
- Test the model on the testing dataset.

Advantages of QSAR:

- Quantifying the relationship between structure and activity provides **an understanding of the effect of structure on activity**.
- It is also **possible to make predictions** leading to the synthesis of novel analogues.
- The results can be used to help **understand interactions** between functional groups in the molecules of greatest activity, with those of their target.

Disadvantages of QSAR

- **False correlations** may arise because biological data that are subject to considerable experimental error (noisy data).
- If training dataset is not large enough, the data collected **may not reflect the complete property space**. Consequently, many QSAR results cannot be used to confidently predict the most likely compounds of best activity.
- Features may not be reliable as well. This is particularly serious for 3D features because 3D structures of ligands binding to receiptor may not be available. Common approach is to use minimized structure, but that may not represent the reality well.

The future of biomolecular nanotechnology on drug discovery

- Generation of improved data processing and managemer infrastructure.
- Introduction of novel research concepts and predictive methods (virtual screening, prediction of molecular transport, prediction of metabolic parameters).
- Computational analysis will reduce the magnitude of experimental programs (compound synthesis and screening).
- Information concerning in vivo assays, pharmacological profiles or in vivo model will be available in databases.