Sequence $\rightarrow$ Structure $\rightarrow$ Function

Alcohol Dehydrogenase

CH$_3$CH$_2$OH $\rightarrow$ H$\overset{\text{O}}{\text{C}}$-CH$_3$

ethanol $\rightarrow$ acetaldehyde
COVID-19

Spike

ACE2 Receptor
If the trial ultimately shows the treatment is effective against Covid-19, it could be available by autumn, according to the Indianapolis-based company.
Primary Structure

Twenty-One Amino Acids

A. Amino Acids with Electrically Charged Side Chains

- **Arginine** (Arg)
  - Positive
  - pKa 12.0

- **Histidine** (His)
  - Positive
  - pKa 6.0

- **Lysine** (Lys)
  - Positive
  - pKa 10.5

- **Aspartic Acid** (Asp)
  - Negative
  - pKa 2.0

- **Glutamic Acid** (Glu)
  - Negative
  - pKa 4.2

B. Amino Acids with Polar Uncharged Side Chains

- **Serine** (Ser)
  - pKa 11.5

- **Threonine** (Thr)
  - pKa 10.8

- **Asparagine** (Asn)
  - pKa 2.2

- **Glutamine** (Gln)
  - pKa 11.8

C. Special Cases

- **Cysteine** (Cys)
  - pKa 11.7

- **Selenocysteine** (Sec)
  - pKa 11.7

- **Glycine** (Gly)
  - pKa 9.5

- **Proline** (Pro)
  - pKa 11.5

D. Amino Acids with Hydrophobic Side Chains

- **Alanine** (Ala)
- **Valine** (Val)
- **Isoleucine** (Ile)
- **Leucine** (Leu)
- **Methionine** (Met)
- **Phenylalanine** (Phe)
- **Tyrosine** (Tyr)
- **Tryptophan** (Trp)

https://en.wikipedia.org/wiki/Amino_acid

N terminal

C terminal
Secondary Structure
Tertiary Structure

TIM barrel fold

beta barrel
Quaternary Structure

domain swapping

hemoglobin
Structure Determination

crystal

x-rays

diffraction pattern

phases

refinement

electron density map

fitting

atomic model

x-ray crystallography

nuclear magnetic resonance

\[ \mathcal{F}\{g(t)\} = G(f) = \int g(t) e^{-2\pi ift} \, dt \]
\[ \mathcal{F}^{-1}\{G(f)\} = \int G(f) e^{2\pi ift} \, df = g(t) \]
Sequence → Structure → Function

Motion
Sequence $\rightarrow$ Structure $\rightarrow$ Function $\uparrow$

Motion
Molecular Dynamics

1. Give atoms initial positions $r^{(t=0)}$, choose short $\Delta t$
2. Get forces $F = - \nabla V(r^{(0)})$ and $a = F/m$
3. Move atoms: $r^{(t+\Delta t)} = r^{(t)} + v^{(t)} \Delta t + \frac{1}{2} a \Delta t^2 + ...$
4. Move time forward: $t = t + \Delta t$
5. Repeat as long as you need

$$V(r^N) = \sum_{\text{bonds}} k_b (l - l_0)^2 + \sum_{\text{angles}} k_a (\theta - \theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_n \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)]$$

$$+ \sum_{j=1}^{N-1} \sum_{i=j+1}^N f_{ij} \left\{ \epsilon_{ij} \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi \epsilon_0 r_{ij}} \right\}$$

https://en.wikibooks.org/wiki/Structural_Biochemistry/Molecular_Modeling/Molecular_Dynamics
Molecular Dynamics

- Give atoms initial positions $r^{(t=0)}$, choose short $\Delta t$
- Get forces $F = -\nabla V(r^{(0)})$ and $a = F/m$
- Move atoms: $r^{(t+\Delta t)} = r^{(t)} + v^{(t)} \Delta t + \frac{1}{2} a \Delta t^2 + ...$
- Move time forward: $t = t + \Delta t$
- Repeat as long as you need

$V(r^N) = \sum_{\text{bonds}} k_b (l - l_0)^2 + \sum_{\text{angles}} k_a (\theta - \theta_0)^2$

$+ \sum_{\text{torsions}} \sum_n \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)]$

$+ \sum_{j=1}^{N-1} \sum_{i=j+1}^N f_{ij} \left\{ \epsilon_{ij} \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$

https://en.wikibooks.org/wiki/Structural_Biochemistry/Molecular_Modeling/Molecular_Dynamics
Dynamics Analysis

More stable

Less stable

ΔRMSF (Bound - Unbound)

p < 0.01
p < 0.001

Residue
Protein Folding

How Fast-Folding Proteins Fold

Kresten Lindorff-Larsen1,*, Stefano Piana1,*, Ron O. Dror1, David E. Shaw1,2,
1D. E. Shaw Research, New York, NY 10036, USA.
2Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.
*To whom correspondence should be addressed. E-mail: david.shaw@DEShawResearch.com (D.E.S.);
kresten.lindorff-larsen@DEShawResearch.com (K.L.-L.); stefano.piana-agostinetti@DEShawResearch.com (S.P.)
†† These authors contributed equally to the manuscript.
See all authors and affiliations

Science 28 Oct 2011;
Vol. 334, Issue 6055, pp. 517-520
DOI:10.1126/science.1208351

https://en.wikipedia.org/wiki/Anton_(computer)
Homology Modeling

Find sequence alignment to template

Build backbone from template (loops??)

Add side-chains

I-TASSER

https://zhanglab.ccmb.med.umich.edu/I-TASSER/
Homology Modeling

Find sequence alignment to template

Build backbone from template (loops??)

Add side-chains

Major source of error: alignment

25% ‘Twilight Zone’
Learning to Predict Structure

https://www.biorxiv.org/content/early/2018/02/14/265231
AlphaFold

Protein Sequence: SQETRKKCTEMKKEFKNCEVRCDESCNSHCVEVRCSDTKYTL

Neural Network → Databases → Distance Predictions → Angle Predictions → Score → Gradient Descent → Structure

Ground truth
Average predicted distance

T0954 / 6CVZ
T0965 / 6D2V
T0955 / 5W9F

https://deepmind.com/blog/alphafold/
AlphaFold

Protein Sequence
SQETRKKCTEMKKFKNCEVRCDESNNHCEVRCSDTYTL

Neural Network
Databases

Distance Predictions
Angle Predictions

Score

Gradient Descent

Structure

Ground truth
Average predicted distance

T0954 / 6CVZ
T0965 / 6D2V
T0955 / 5W9F

https://deepmind.com/blog/alphafold/
Computational Drug Discovery
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

POTENTIAL NEW MEDICINES

- BASIC RESEARCH
- DRUG DISCOVERY
- PRE-CLINICAL
- CLINICAL TRIALS
- FDA REVIEW
- POST-APPROVAL RESEARCH & MONITORING

IND SUBMITTED
NDA/BLA SUBMITTED
FDA APPROVAL
TENS
HUNDREDS
THOUSANDS
NUMBER OF VOLUNTEERS

PHASE I
PHASE II
PHASE III
PHASE IV

1 FDA-APPROVED MEDICINE

$2.6 BILLION

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
Drug Discovery

Omics
Target Identification

Compounds
Screening

Hits
Lead Identification

Leads
Lead Optimization

Clinical Candidates

Cost
Computational Drug Discovery

Omics

Target Identification

Compounds

Virtual

Screening

Hits

Lead Identification

Leads

Cost

Clinical Candidates

Modeling

Lead Optimization
Ligand Based Drug Design

Fingerprint Methods
- map molecules to a descriptor space:
  1D: molecule weight, #h-bonds, etc.
  2D: paths, bond distances between atom-pairs
  Example: Daylight/FP2
    • all paths up to 7 bonds long
    • each path corresponds to a bit
- similarity is “distance” between descriptors
- for bit vectors, Tanimoto distance used

\[ T(A, B) = \frac{|A \cap B|}{|A \cup B|} \]
Ligand Based Drug Design

Fingerprint Methods
- map molecules to a descriptor space:
  1D: molecule weight, #h-bonds, etc.
  2D: paths, bond distances between atom-pairs
Example: Daylight/FP2
  • all paths up to 7 bonds long
  • each path corresponds to a bit
- similarity is “distance” between descriptors
- for bit vectors, Tanimoto distance used

\[ T(A, B) = \frac{|A \cap B|}{|A \cup B|} \]
Ligand Based: QSAR

Quantitative Structure/Activity Relationships

Regression model used to estimate relationship between variables

Model identifies relationship between 2D chemical structures and bioactivity

The vectors correspond to the bits

Weights are assigned based on prevalence of substructures

A relationship between vectors, or substructures, and a property can be estimated

\[ f(\vec{x}) = w_1 \vec{x}_1 + w_2 \vec{x}_2 + w_3 \vec{x}_3 + \ldots + b \]
Structure Based Drug Design

Unlike ligand based approaches, generalizes to new targets.

Requires **molecular target** with **known structure** and **binding site**.
Structure Based Drug Design

Virtual Screening

Lead Optimization

Pose Prediction

Binding Discrimination

Affinity Prediction
Structure Based Drug Design

Virtual Screening

Lead Optimization

Pose Prediction

Binding Discrimination

Affinity Prediction
Protein-Ligand Scoring

AutoDock Vina

\[
\text{gauss}_1(d) = \frac{w_{\text{gauss}_1}}{(d/0.5)^2}
\]

\[
\text{gauss}_2(d) = \frac{w_{\text{gauss}_2}}{(d-3)^2}
\]

\[
\text{repulsion}(d) = \begin{cases} 
  w_{\text{repulsion}}d^2 & d < 0 \\
  0 & d \geq 0 
\end{cases}
\]

\[
\text{hydrophobic}(d) = \begin{cases} 
  w_{\text{hydrophobic}} & d < 0.5 \\
  0 & d > 1.5 \\
  w_{\text{hydrophobic}}(1.5 - d) & \text{otherwise}
\end{cases}
\]

\[
\text{hbond}(d) = \begin{cases} 
  w_{\text{hbond}} & d < -0.7 \\
  0 & d > 0 \\
  w_{\text{hbond}}(-\frac{10}{7}d) & \text{otherwise}
\end{cases}
\]

Pharmacophore

IUPAC: The ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response.

**Common Features:**
- aromatic ring
- hydrophobic area
- positive ionizable
- negative ionizable
- hydrogen bond donor
- hydrogen bond acceptor
Compound 2.12 vs hcSHMT

\[ K_{\text{app}} = 24 \] PM

\[ \text{Abs} \]

Tim

Actin

MIV

MSV

\[ \text{PM} \]
Purchasable

Accessible

Matching

Scoring

Dynamics

Compound 2.12 vs hcSHMT

$K_i = 24$

$P_M$ [Inhibitor] ($P_M')$

Abs Tim

Actin

MIV

MSV
http://pharmit.csb.pitt.edu

4PPS

DUDe ER alpha benchmark
Metrics

- **sensitivity, recall, hit rate, or true positive rate (TPR)**
  \[
  \text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR}
  \]

- **specificity, selectivity or true negative rate (TNR)**
  \[
  \text{TNR} = \frac{\text{TN}}{N} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR}
  \]

- **precision or positive predictive value (PPV)**
  \[
  \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}
  \]

- **accuracy (ACC)**
  \[
  \text{ACC} = \frac{\text{TP} + \text{TN}}{P + N} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
  \]

- **F1 score**
  is the harmonic mean of precision and sensitivity
  \[
  F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}
  \]

https://en.wikipedia.org/wiki/F1_score