A Quantitative Universe

Brian Y. Chen
Lehigh University
Dept. Computer Science & Engineering
My First Computer: Apple ][ Plus, 1981
Informatics is the science of designing methodologies for gathering, analyzing, integrating, and visualizing data used to form an opinion.
Translation Off

Translation On

Word Lens, iPhone App released: Dec 16, 2010
GenBank – Global public repository of DNA sequences

Protein Data Bank – Global public repository of protein structures
What is Bioinformatics?

The science of designing methodologies for gathering, analyzing, integrating, and visualizing data used to form an opinion on Biological Systems.
Biological systems are nested and interacting machines
Individual molecules are the foundation of all systems
Structural biology studies molecular machines
Structural biology connects structure with function
Structural biology has become a quantifiable universe.

Timeline of Nobel Prizes in Structural Biology:
- 1946: Sumner
- 1962: Crick, Watson, Wilkins
- 1962: Perutz, Kendrew
- 1964: Hodgkin
- 1972: Anfinsen
- 1982: Klug
- 1988: Deisenhofer, Huber, Michel
- 1991: Ernst
- 1997: Walker
- 2002: Wuthrich
- 2003: MacKinnon
- 2006: Kornberg
- 2009: Steitz, Yonath
- 2009: Ramakrishnan
- 2009: Steitz, Yonath

Number of Entries in the Protein Data Bank:
- Source: www.pdb.org
Structural bioinformatics connects structure with function at scale and with precision.
Structural bioinformatics draws from many computational fields.
The General Problem:

Gather, analyze, integrate, and visualize data used to form an opinion on Biological Systems.
Proteins are chains of amino acids
Similar sequences imply similar function

HAWPFMVMLQL-AGG------HFCGATLIAIPNFVMSAAHCVCANVNV
HAWPFMVMLQL-RGG------HFCGATLIAIPNFVMSAAHCVCANVK-
HSWPWQISLQY-SKNDAWGHTCGGTIASNYVLTAHCISNAKT
HSRPYMVSLQV-Q---G-NHFCGGTLIHPQFVMTAAHCIDKINP
LA-PYIASLQRN-RGG------HFCGGTLIHQQFVMTAAHCINSRNV

Similar functional sites imply similar function

Motif

Known function

Match

Target

Unknown function

MASH

Combinatorial Extension
Nussinov et al, Proteins, 2001

Geometric Hashing

pevoSOAR

Ska

Geometric Sieving

PINTS

JESS
Holm et al, Bioinformatics, 2008.
Protein surfaces reveal functional sites

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
</table>
Similarity doesn’t tell us everything

How does this protein fit in the system?

What parts of the protein make it work?
Specificity is preferential binding

Specificity is an aspect of function
Cavity shape influences specificity
Proteins with the same function can have different specificity
VASP isolates differences in cavity shape

VASP: Volumetric Analysis of the Surfaces of Proteins
• Identify amino acids that alter cavity shape
• Identify subcavities that alter cavity shape

Results: VASP finds influences on specificity
The VASP procedure

Input: A Protein Family

Align Structures

Define Cavities

Volumetric Comparison Of Cavities

Output: Volumetric Differences
The VASP procedure

Input: A Protein Family
Align Structures
Define Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences

The VASP procedure

- Input: A Protein Family
- Align Structures
- Define Cavities
- Volumetric Comparison Of Cavities
- Output: Volumetric Differences

Computational Solid Geometry
Computational Solid Geometry (CSG)
CSG was originally for modeling parts
Computational Solid Geometry (CSG)

- Union
- Intersection
- Difference
Using CSG with protein structures
The VASP procedure

- Input: A Protein Family
- Align Structures
- Define Cavities
- Volumetric Comparison Of Cavities
- Output: Volumetric Differences
Begin with the molecular surface
Compute an envelope surface
Find the interior surface
Identify nearby amino acids
Compute the convex hull

CSG hull minus molecular surface
CSG intersection with the envelope surface
Remove disconnected pieces
The VASP procedure

Input: A Protein Family
Align Structures
Define Functional Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences

- Amino Acids affecting cavity shape
- Subcavities affecting cavity shape
Finding amino acids that affect cavity shape
Finding amino acids that affect cavity shape
Finding amino acids that affect cavity shape
The VASP procedure

Input: A Protein Family
Align Structures
Define Functional Cavities
Volumetric Comparison Of Cavities
Output: Volumetric Differences

• Amino Acids affecting cavity shape
• Subcavities affecting cavity shape
What makes A cavities different from B?
What is common in A?
What is the maximum extent of B?
All parts of A that are not in any part of B

Intersection: A \cap B

Union: A \cup B

Difference: A \setminus B

Output: A \setminus B
Results

• Serine Proteases: Same function, different specificity
  – Trypsins
  – Elastases
  – Chymotrypsins

• Experiments
  – VASP identifies amino acids that influence specificity
  – VASP identifies subcavities that influence specificity
The serine protease family

Serine Proteases

Chymotrypsin Clan
Catalytic Triad: His-Asp-Ser
- Chymotrypsins

Subtilisin Clan (Asp-His-Ser)
- Subtilisins

Other clans (not used)
- Oligopeptidases (Asp-Ser-His)
- Carboxypeptidases (Ser-Asp-His)
- Others...

Serine proteases break up other proteins

A serine protease up close
Structural Alignment
Alignment by Catalytic triad + S1 residue
(Cα and Cβ atoms)

Chymotrypsins
3.4.21.1

Trypsins
3.4.21.4

Elastases
3.4.21.36

Alignments for different enzymes.
Serine proteases have specificity for different sequences of amino acids.

Chymotrypsins prefer big amino acids

Ser 189

Triad

{ Tyr, Phe, Trp }
Trypsins bind positively charged residues

Elastases prefer small amino acids

Elastases

Porcine Pancreatic Elastase 1b0e

{ Ala, Gly, Val, .. }

P1 P2 P1' C

Ser 189

Triad
The data is filtered for noise and bias

- **Chymotrypsins**
  - 58 individual structures
  - 45 Non-mutants
  - 2 < 90% Sequence Identity

- **Trypsins**
  - 371 individual structures
  - 290 Non-mutants
  - 11 < 90% Sequence Identity

- **Elastases**
  - 91 individual structures
  - 91 Non-mutants
  - 2 < 90% Sequence Identity
VASP finds amino acids in elastase that influence specificity

VASP finds amino acids in trypsins that influence specificity

VASP finds subcavities in trypsins and elastases that influence specificity.
Discussion

- **VASP can identify:**
  - Amino acids that influence specificity
  - Subcavities that influence specificity

- **Contributions**
  - The first unsupervised analysis of protein structures that identifies active components of functional sites
  - The first algorithm to isolate the basis for specificity in protein structures
  - The first representation of proteins using smooth solid volumes

- **What can we use VASP for?**
  - Identify amino acids that might change specificity in drug resistance
  - Influential subcavities point to drug designs that bind more specifically, and thus reduce side effects
Specificity is important in all systems
VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity

Brian Chen and Barry Honig
_PLOS Computational Biology._ 6(8): e1000881. doi:10.1371/journal.pcbi.1000881. (in print as of August 12, 2010)
VASP is only the beginning
CSE 308/408
Bioinformatics: Issues and Algorithms

• Concentration on genomics and bioinformatics
• Learn about how DNA sequencing is transforming biology and medicine
• Topics:
  • Genome Assembly
  • Genome Annotation
  • Genome Evolution
• Project options include projects that can be completed without programming (Enterprise computer usage is necessary though)
Acknowledgements

- **Honig Lab**
  - Barry Honig
  - Markus Fischer
  - Donald Petrey
  - Remo Rohs
  - Jeremie Vendome
  - Fabian Dey
  - Shana Posy

- **Murray Lab**
  - Diana Murray
  - Paul Murray
  - Hunjoong Li
  - Janey Zhang
  - Tonya Silkov

- **Kavraki Group**
  - Lydia Kavraki
  - Drew Bryant
  - Amanda Cruess
  - Brad Dodson
  - Joey Bylund

- **Kimmel Group**
  - Marek Kimmel
  - Viacheslav Y. Fofanov

- **Lichtarge Lab**
  - Olivier Lichtarge
  - David M. Kristensen

- **Linusson Lab**
  - Anna Linusson
  - David Andersson

- **Funding (PhD)**
  - NSF
  - NLM, NIGMS
  - HHMI
  - Whitaker Found.
  - Sloan Found.
  - AMD

- **Organizations**
  - C₂B²
  - Keck Center

- **Fumiko Chino**
Questions