Perth Computational Biology Workshop June 2004

Bioinformatics II - Evolution of Protein Structure
Evolution of Protein Structure in the Aminoacyl-tRNA Synthetases

Zan Luthey-Schulten
Department of Chemistry, Beckman Institute, Center for Biophysics and Computational Biology
University of Illinois at Urbana-Champaign
Aminoacyl-tRNA synthetases

Universal Tree of Life

Structural Conservation in the Catalytic Domain of the AARSs

Class I Lysyl-tRNA Synthetase

Class II Lysyl-tRNA Synthetase
Why Study the Evolution of Protein Structure?

1. Important for Homology Modeling
   Better profiles improve database searches and give better alignments of distant homologs. Allows mixing of sequence and structure information systematically.

2. Learn how evolutionary dynamics changed protein shape.
   Mapping a protein of unknown structure onto a homologous protein of known structure is equivalent to defining the evolutionary pathway connecting the two proteins.

3. Impact on protein structure prediction, folding, and function
   Evolutionary profiles increase the signal to noise ratio.
Outline


Methods

2. Introduce a structure-based metric which accounts for gaps, and show that evolutionary information is encoded in protein structure.

3. Introduce multidimensional QR factorization for computing non-redundant representative multiple alignments in sequence or structure.

Applications

4. Non-redundant multiple alignments which well represent the evolutionary history of a protein group provide better profiles for database searching.

   Eliminate bias inherited from structure or sequence databases.

   Important for bioinformatic analysis (substitution matrices, knowledge based potentials structure pred., genome annotation) and evolutionary analysis.

5. Depict the evolution of structure and function in Aspartyl-tRNA synthetase.
Universal Phylogenetic Tree
three domains of life

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese PNAS 2000

Woese, Olsen, Ibba, Soll MMBR 2000
Phylogenetic Distributions

Full Canonical  Basal Canonical  Non-canonical

increasing inter-domain of life Horizontal Gene Transfer

“HGT erodes the historical trace, but does not completely erase it….” G. Olsen
Protein Structure Similarity Measure

**$Q_H$ - Structural Homology**
fraction of native contacts for aligned residues + presence and perturbation of gaps

$$Q_H = N \left[ q_{aln} + q_{gap} \right]$$

$$q_{aln} = \sum_{i < j - 2} \exp \left[ -\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2} \right]$$

“Gaps should count as a character but not dominate” C. Woese

O’Donoghue & Luthey-Schulten *MMBR.2003.*
Protein structure encodes evolutionary information

**Sequence Phylogeny**
AspRS-AsnRS Group

**Structure Phylogeny**
AspRS-AsnRS Group

Woese, Olsen, Ibba, Soll *MMBR* 2000

Horizontal Gene Transfer in Protein Structure

Sequence Phylogeny
AspRS-AsnRS Group
Non-redundant Representative Sets

Multidimensional QR factorization of alignment matrix, $A$.

$A = \begin{bmatrix} X & Y & G \\ Z & l_{aln} & k_{proteins} \end{bmatrix}$

QR computes a set of minimal linearly dependent structures.

Numerical Encoding of Proteins in a Multiple Alignment

**Encoding Structure**
Rotated Cartesian + Gap = 4-space

- **Aligned position**
  \( (x_C, y_C, z_C, 0) \)

- **Gapped position**
  \( (0, 0, 0, g) \)

- **Gap Scaling**
  \[ g = \frac{\|X\|_F + \|Y\|_F + \|Z\|_F}{\|G\|_F} \]

- Adjustable parameter

**Sequence Space**
Orthogonal Encoding = 24-space

- 23 amino acids (20 + B, X, Z) + gap

  - \( A = (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \)
  - \( B = (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \)
  - \( C = (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \)
  - ...\n  - \( GAP = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1) \)

**Alignment Matrix**

- \( A = \) encoded residue space
- \( m \) aligned positions
- \( n \) proteins
- \( d = 1, 2, 3, \ldots, N \)
A Multiple Alignment is a Matrix with Linearly Dependent Columns

redudancy is equivalent to linear dependence

**QR factorization**

Re-orders the columns of $A$, segregating the linearly independent columns from the dependent ones without scrambling the information in $A$. SVD not an option.

$$Q^T A P = \tilde{R}$$

$$\tilde{A} = A P$$

$Q^T$ – orthogonal matrix of product of Householder transformations.

$P$ – permutation matrix encodes column pivoting which exchanges columns of $A$ and puts the redundant or similar proteins to the right hand side.

**Multidimensional QR**

$N$ simultaneous QR factorizations, one for each $d$-dimension.

A minimal linearly dependent subset can be determined with respect to a threshold, e.g., similarity measure threshold.


The QR establishes an **order** of linear dependence by applying Householder transformations and permutations

\[ Q^T = H_n...H_1 \]

Three 1-D (2 residue) proteins \( \mathbf{a} \ \mathbf{b} \ \mathbf{c} \).

\( \mathbf{a} \) is our measuring stick, reference frame.

The transformation reveals that \( \mathbf{b} \) is more linearly dependent on \( \mathbf{a} \), so the permutation swaps \( \mathbf{b}' \) with \( \mathbf{c}' \).

Given \( \mathbf{a}, \mathbf{c} \) adds more information to the system than \( \mathbf{b} \).

Multiply aligned proteins exist in a higher dimensional space, so this magnitude is computed with a matrix \( p \)-norm:

\[
\| \mathbf{a}_j \|_{F_p} = (\sum_{d=1}^{4} \sum_{i=k}^{m_{d,n}} |a_{i,j,d}|^{p})^{\frac{1}{p}}
\]

Class I AARSs

evolutionary events

5 Subclasses

Specificity – 11 Amino acids

Domain of life A, B, E
How many sequences are needed to represent the Subclass ILMV?

If each of ILMV was full canonical, then we would need $4 \times 3 = 12$ sequences.

Since M and V are basal, we need at least $2 \times 3 + 2 \times 2 = 10$ sequences.

We have 6 structures.
Non-Redundant Profiles for Database Searching
AARS Subclass ILMV

Starting with a non-redundant profile, accuracy diminishes with Psi-blast iterations which add in bias. Repair with QR filter.

Choosing the right 10 sequence makes all the difference.

A. Sethi, P. O’Donoghue, Z. Luthey-Schulten
Evolutionary Structure/Sequence Profiles Suggest Reaction Pathway

Summary

Evolutionary information is encoded in protein structure.

- Protein structure can be used to investigate early evolutionary events.
- Accounting for gaps is important for comparing homologous structures.

Multidimensional QR factorization computes non-redundant sets from multiple sequence or structure alignments which well represent the evolutionary history of the group.

Structure databases are limited, but multiple structural alignments provide accurate alignments, especially in the case of distant homologies.

Supplement the structures with an appropriate number and type of sequences (in accord with the phylogenetic topology) to produce minimal representative profiles.
Acknowledgements

Funding: NSF, NIH, NIH Resource for Macromolecular Modeling and Bioinformatics

Collaborators

Evolutionary Studies
Gary Olsen, Carl Woese (UIUC)

Algorithms
Mike Heath (UIUC)
Rob Russell (EMBL) STAMP

Protein Structure Prediction
Peter Wolynes, Jose Onuchic, Ken Suslick

Graphics Programmers VMD
John Stone, Dan Wright, John Eargle

Patrick O’Donoghue
Rommie Amaro
Anurag Sethi
John Eargle
Corey Hardin
Michael Baym
Michael Januszyk

Felix Autenrieth
Taras Pogorelov