Evolution of Protein Structure in the Aminoacyl-tRNA Synthetases

P. O’Donoghue and Z. Luthey-Schulten*
Department of Chemistry, Beckman Institute, Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign
What can be learned from AARS?

• “The aminoacyl-tRNA synthetases, perhaps better than any other molecules in the cell, epitomize the current situation and help to under standard (the effects) of HGT” Woese (PNAS, 2000; MMBR 2000)
Aminoacyl-tRNA synthetases

Universal Tree of Life

Bacteria

Archaea

Eucarya

Crenarchaeota

Euryarchaeota

Fungi

Plants

Animals

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 - Evolution is the foundation of bioinformatics.
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

Archaea
Eucarya
Bacteria

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese *PNAS* 2000

Woese, Olsen, Ibba, Soll *MMBR* 2000
After W. Doolittle, modified by G. Olsen
Phylogenetic Distributions

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

$$Q_H = \prod [q_{aln} + q_{gap}]$$

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

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

Protein structure encodes evolutionary information

**Sequence Phylogeny**
AspRS-AsnRS Group

**Structure Phylogeny**
AspRS-AsnRS Group
Class II AARSs

Woese, Olsen, Ibba, Soll *MMBR* 2000

Horizontal Gene Transfer in Protein Structure

Sequence Phylogeny
AspRS-AsnRS Group
Non-redundant Representative Sets

Too much information
129 Structures

Multidimensional QR factorization of alignment matrix, $A$.

Economy of information
16 representatives

$$A = \begin{bmatrix} l_{x} & l_{y} & l_{z} & d=4 \\ X & Y & Z & G \end{bmatrix}$$

QR computes a set of maximal linearly independent structures.

Numerical Encoding of Proteins in a Multiple Alignment

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

**Aligned position**  
$(x_{C\alpha}, y_{C\alpha}, z_{C\alpha}, 0)$

**Gapped position**  
$(0, 0, 0, g)$

**Gap Scaling**  
$$g = \frac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}$$

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

23 amino acids  $(20 + B, X, Z) + \text{gap}$

- A = $(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- B = $(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- C = $(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- GAP = $(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$

**Alignment Matrix**
A Multiple Alignment is a Matrix with Linearly Dependent Columns

redundancy 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

$$Q^T = H_n \ldots H_1$$

Three 1-D (2 residue) proteins **a b c**.

**a** is our measuring stick, reference frame.

The transformation reveals that **b** is more linearly dependent on **a**, so the permutation swaps **b'** with **c'**.

Given **a, c** adds more information to the system than **b**.

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

$$\|a_j\|_{F_p} = \left( \sum_{d=1}^{4} \sum_{i=k}^{m_{a1n}} |a_{i,j,d}|^p \right)^{1/p}$$

What are the constraints on the parameters?
Must maintain the evolutionary history of the protein group.

This rule is used to determine the value of two adjustable parameters in our implementation of the QR.
Hierarchical Multidimensional QR -
Parameters Define the Measure of Linear Dependence

AARS class I, Rossman fold

AARS class II, Novel fold

ordering norm

\[ \gamma \text{ (normalized)} \]

\[ \max_{j=k,\ldots,n_{\text{proteins}}} (\|a_j\|_{F_p}) \]

\[ \|a_j\|_{F_p} = \left(\sum_{d=1}^{4} \sum_{i=k}^{m_{\text{ain}}} |a_{ijd}|\right)^{1/p} \]

ordering p-norm

allowed

forbidden

gap scale

\[ g = \frac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}} \]
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.

A. Sethi, P. O’Donoghue, Z..Luthey-Schulten
Combined Structure-Sequence Phylogeny
an evolutionary profile of subclass IA AARSs

\( \delta Q_H = 0.1 \)
\( \delta S_i d = 0.25 \)

Pfam profile composition:
- Le \( x 0 \)
- La \( x 0 \)
- Lb \( x 3 \)
- La \( x 4 \)
- Ia \( x 1 \)
- Ile \( x 4 \)
- Vb \( x 7 \)
- Va \( x 0 \)
- Ma \( x 1 \)
- Mb \( x 0 \)

Number of false positives vs. Number of true positives:
- 10 sequences
- 20 sequences
- 6 sequences
- 113 sequences
- 28 sequences
- 16 sequences

Profiles:
- QR structure
- PFAM
- QR str. + 4 seqs
- QR str. + 12 seqs
The Economy of Information
How many sequence are needed for profiles?

A single profile for class I AARSs

PFAM profile of 113 sequences finds 3 additional sequence fragments compared to the non-redundant profile of 28 sequences.

HisA and HisF Protein Family
TIM Barrel fold

If the sequences well represent the evolutionary history of the protein family, a factor of 10 to 100 less information is required.
Evolutionary Structure/Sequence Profiles Suggest Reaction Pathway

Evolution of Structure and Function in AspRS
AARS domains have different Evolutionary Histories

catalytic domains
AARSs II

bacterial type aspartyl-tRNA synthetase
*E. coli*, homodimer
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 - structure metric

Multidimensional QR factorization computes non-redundant sets from multiple sequence or structure alignments which well represent the evolutionary history of the group as expressed in phylogenetic tree

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. Search profiles for foldons!!
Evolution of Protein Structure
VMD Multiple Sequence Display with Evolution Analysis Algorithms
Acknowledgements

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

Felix Autenrieth
Taras Pogorelov
Brijeet Dhaliwal

Graphics Programmers VMD
John Stone, Dan Wright, John Eargle
http://www.ks.uiuc.edu/Research/vmd/alpha/zs04/

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

Funding: NSF, NIH, NIH Resource for Macromolecular Modeling and Bioinformatics, NRAC NSF Supercomputer Centers
extra slides
Structure Phylogeny
Class I AARSs

Structure Phylogeny
Class II AARSs
Structural Overlap of the AARSs

Class I

Class II

A

D

B

E

C

F
Structural Conservation in tRNA

T-loop

acceptor stem

anticodon loop

D-loop
Representative set of OB folds involved in translation
Only structure can reveal distant evolutionary relationships
Conservation of Sequence and Structure

GlnRS

AsnRS
Protein structure encodes evolutionary information

Sequence Phylogeny
Woese et al. 2000

Structural Overlap

T. thermophilus
P. kodakaraensis

Structure Phylogeny
Class II AARSs
Horizontal Gene Transfer and Protein Structure in ProRS

Sequence Phylogeny
Woese et al. 2000

Structure Phylogeny
Class II AARSs

Structural Overlap

T. thermophilus
M. thermoautotrophicus
Structural Homology Measure
the effect of insertions

Q_H = 0.82

0.70

0.62

“A gaps influence the analysis
But should not dominate it” CW
Structural Homology Measure
compare inserted residues to gap edges

\[ Q_H = \mathbb{K} \left[ q_{aln} + q_{gap} \right] \]

\[ q_{gap} = \sum_{g_a} \sum_{j} \max \left\{ \exp \left[ -\frac{(r_{g_a,j} - r_{g'_a,j'})^2}{2\sigma_{g_a,j}^2} \right], \exp \left[ -\frac{(r_{g_a,j} - r_{g''_a,j'})^2}{2\sigma_{g_a,j}^2} \right] \right\} \]

\[ + \sum_{g_b} \sum_{j} \max \left\{ \exp \left[ -\frac{(r_{g_b,j} - r_{g'_b,j'})^2}{2\sigma_{g_b,j}^2} \right], \exp \left[ -\frac{(r_{g_b,j} - r_{g''_b,j'})^2}{2\sigma_{g_b,j}^2} \right] \right\} \]
QR Factorization

Solve the least squares problem \( Ax = b \)

by triangularizing \( A \) with an orthogonal transformation.

\[
Q^T A = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad Q^T(Ax) = Q^T(b)
\]

The system is now solved by back substitution,

\[
\begin{bmatrix} R \\ 0 \end{bmatrix} x = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \quad Rx = c_1
\]

with a minimum residual of

\[
\|r\|_2 = \|b - Ax\|_2 = \|c_2\|_2
\]
Multi-Dimensional QR

N-dimensional QR = N one-dimensional QRs.

Permutation matrix is constant for each dimension, ordering norm is Frobenius-like matrix p-norm.

$$\max_{j=k,\ldots,n_{\text{proteins}}} (\|a_j\|_{F_p}) \quad \|a_j\|_{F_p} = \left(\sum_{d=1}^{4} \sum_{i=k}^{m_{\text{alin}}} |a_{ijd}|^p\right)^{1/p}$$

Encoding Structure

Aligned residues: $$(x_{C\alpha}, y_{C\alpha}, z_{C\alpha}, 0)$$

Gap “residues”: $$(0, 0, 0, g)$$

Gap Scaling

$$g = \gamma \frac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}$$

Encoding Sequence

Orthogonal Encoding = 24-space

23 amino acids symbols (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)
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)
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)

...  
GAP=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)

QR Factorization with Column Pivoting

1. Calculate column norm of column $i$ and all columns to the right.

$$A^{(k-1)} = H_{k-1}...H_1AP_1...P_{k-1}$$

$$\max_{j=k,...,n} (s_j^{(k)}) \quad s_j^{(k)} = \left( \sum_{i=k}^{m} a_{ij}^2 \right)^{1/2}$$

2. Swap column $i$ with column to the right of maximum norm and record column permutation.

$$H_{k-1}...H_1AP_1...P_{k-1}P_k$$

3. Construct and apply $H_k$

$$A^{(k)} = H_k H_{k-1}...H_1AP_1...P_{k-1}P_k$$

$$\tilde{A} = AP_1...P_n = AP$$

Original matrix, $A$, columns ordered by increasing linear dependence.

Golub, *Numerische Mathematik*, 1965
Protein Structure Prediction

Ab Initio protein folding

\[ E_{AM} = - \sum_{\mu=1}^{N} \sum_{i,j} \{ \gamma_{AM} [ P_i, P_j, P_i^\mu, P_j^\mu ] \} \]

\[ X \exp \left[ \frac{-(r_{ij} - r_{ij}^\mu)^2}{2\sigma_{ij}^2} \right] \]

1-D protein sequence

SISSIRVKSRIQLG...

Threading/Profile Alignment

\[ E = E_{match} + E_{gap} \]

Target Sequence

SISSIRVKSRIQLGLNQAELA-QKV------GTTQ...

QFANEFKVRRIKLGYTQ------TNVGEALAAVHG...

Known structure(s)

3-D protein structure

Eastwood, Hardin, Luthey-Schulten, Wolynes (2001)
IBM. J. RES. & DEV. 45:475-497

Sequence-Structure Alignment

Target sequence

A_1 A_2 A_3 A_4 A_5 ... 

Alignment between target(s) and scaffold(s)

1. Energy Based Threading*

\[ H = E_{contact} + E_{profile} + E_{H-bonds} + E_{gap} \]

\[ E_{profile} = \sum_{i}^{n} \gamma^{(p)}(A_i, SS_i, SA_i) \]

\[ E_{contact} = \sum_{i,j}^{2} \sum_{k=1}^{\gamma^{(ct)}(A_i, A_j)} U(r_k - r_ij) \]

2. Sequence – Structure Profile Alignments

Clustal, Hidden Markov (HMMER, PSSM)
with position dependent gap penalties

The prediction is never better than the scaffold.

Threading energy function/profiles requires improvement.
Why Study the Evolution of Protein Structure?

In what specific ways has the evolutionary dynamic changed protein shape over time?

What can studying the change in protein shape over time tell us about the evolutionary process?

How did translation evolve?

When, with respect to the root of the universal phylogenetic tree, was translation established in its modern form?

What was the role of the AARSs in the evolution of the translation mechanism, development of the genetic code?