Adaptive protein evolution: Introduction

Belinda Chang
Department of Ecology & Evolutionary Biology
Department of Cell & Systems Biology
University of Toronto
@opsinlab

How do protein sequences evolve?

Can we identify evolutionary patterns of selection associated with adaptive shifts in protein function?

Can we identify the underlying mechanisms associated with adaptive shifts?
Evolution of protein function

\[ Q_i = \begin{bmatrix} 0 & \pi_i & \omega \pi_i & \omega \kappa \pi_i \end{bmatrix} \]

Rapid accumulation of sequence data
Comparative sequencing can be used to address questions at many different levels

- Evolution of organisms, systematics
- Evolution of genomes
- Evolution of gene regulation
- Evolution of proteins

Evolution of protein-coding genes: Comparative sequence analysis

- Phylogenetically based methods
- Models of evolution (nucleotide, amino acid, codon)
- Hypothesis testing of theories of selection
- $dN/dS$ as a measure of the strength of selection
Increased use of codon models

Codon-based testing for positive selection: Why so popular? Hypothesis testing!

- WHEN selection occurred in evolution
  - Episodic, pervasive, lineage-specific selection
- WHICH proteins were targets of selection
  - Physiology: sensory, metabolic, developmental
- WHICH regions of the protein
  - Mechanisms underlying evolution of function
- Hints as to WHY selection occurred
Evolution of protein function

\[
Q_i = \begin{cases} 
0 & \text{if } i = j \\
\pi_i & \text{and } j \text{ differs by } >1 \\
\kappa_i & \text{if } j \text{ for synonyms} \\
\omega_i & \text{if } j \text{ for non-synonyms} \\
\omega \kappa_i & \text{for non-synonyms}
\end{cases}
\]

Rapid accumulation of protein structures

- Parameter estimation via ML
- Stationary process
- Phenomenological parameters
- ts/tv ratio
- Codon frequencies
- \( \omega = \frac{dN}{dS} \)
Rapid accumulation of protein structures

- Driven by interest in high-throughput crystallography
- Advances in protein structure determination methods
- Programs such as the Protein Structure Initiative
- Targeted difficult to crystallize proteins such as membrane proteins and large macromolecular assemblies

G protein-coupled receptors

- Largest family of TM signaling proteins
- Extremely difficult to crystallize
Conformational changes upon activation

Chloe et al 2011 Nature

Largest monomeric TM protein: FecA

- Bacterial ion transporter
- Large transmembrane protein
- 22 beta strands
Largest monomeric TM protein: FecA

- Bacterial ion transporter
- Large transmembrane protein
- 22 beta strands

Nuclear pore complex

- Largest membrane bound structure
- About 30 different proteins
- Diameter of 98 nm, 50 MDa
Recent advances in protein structure studies

- Difficulties of working with proteins
- Required the development of expression methods to obtain large amounts of properly folded protein
- Mostly X-ray crystallography, but also NMR, and more recently cryo-electron microscopy
- Homology modeling, molecular dynamics
- Structure predictions

In vitro protein expression methodologies

- Grow cells containing gene of interest
- Purify protein
- Functional assays
- Protein structure
In vitro expression vs. purification from tissue

- Many proteins only present in small amounts in tissue
- Purity of sample may be an issue with complex tissues
- Purification from tissue samples does not allow for site-directed mutagenesis studies
- In vitro expression allows for testing of evolutionary hypotheses of protein structure and function

In vitro protein expression methodologies

- Bacteria: E. coli
- Yeast cells: S. cerevisiae
- Insect cells: SF9
- Mammalian cell culture: HEK293
**Protein structure methodologies**

**X-ray crystallography**

- Multiple conformations, flexible regions often unresolved
- Crystallization conditions not found in nature
- Serial femtosecond crystallography

---

**Protein structure methodologies**

**NMR spectroscopy**

- Advantage of measuring proteins in solution
- Great for studying flexible proteins
- Limited to small proteins
Advances in direct detection, sample prep, and instrumentation have achieved high resolution for larger protein complexes. This technique offers high resolution of larger proteins in a native state. Requires highly specialized instrumentation.

What about protein structure prediction?

Where the twilight zone is may depend on your view.

Krieger et al. 2003
Protein structure prediction

- Homology modeling
  - MODELLER (https://salilab.org/modeller/)
  - Rosetta suite (http://robetta.bakerlab.org/)
  - SWISS-MODEL (https://swissmodel.expasy.org/)

- Molecular dynamics simulations

The twilight zone for ab initio protein folding predictions

Dill & MacCallum, Science 2012
Molecular evolution: Evolution of protein function

DATA
- Genomic sequencing
- Protein structures

TOOLS
- Phylogenetic models of coding sequence evolution
- Experimental studies, mutagenesis and ancestral resurrection

Phylogenetic approaches to the study of protein structure and function

- Ancestral protein reconstruction
- Computational analyses of selection (dN/dS)

-> Combining computational with experimental approaches allows us to test hypotheses of selection in protein evolution
Resurrecting ancestral proteins

Thornton, 2004
Nat. Reviews Genet. (5):366

Ancestral reconstruction: considerations

- Most studies use ML/Bayes methods to infer ancestral sequence with highest probability, single point estimate
- Violations of model assumptions, e.g. shifts in equilibrium frequencies
- Uncertainty in tree topology
- Statistical bias towards states with highest equilibrium frequencies
- This may also result in functional bias towards more stable proteins (Goldstein et al. 2013)

-> How to assess robustness of reconstruction in a functional context?
Assessing robustness of reconstruction in a functional context

- Alternate tree topologies, species tree topology
- Alternate approaches, models of evolution
- Sampling alternate ancestors from the posterior distribution (Pollack & Chang 2012)
- Sampling of near ancestor sequences (Bar-Rogovsky et al. 2015)
- Uncertainty in genotype does not necessarily reflect uncertainty in phenotype (Gaucher et al. 2008)

-> Need for experimental data to inform effects of uncertainty in reconstruction on function

Uricase evolution in primates

• Key enzyme metabolizing uric acid in vertebrates
• Lost in some primates, including humans
• Prevalence of diseases such as gout, hypertension, obesity, cardiovascular disease
• Uricase knockouts in mice result in mortality in first 4 weeks

-> Kratzer et al. 2014 (PNAS) used experimentally recreated ancient uricases to determine exactly when, and how, uricase function was lost in primates.
enzymatically abolished before the pseudogenization events in activity of An30 demonstrates that uricase function was nearly significantly more catalytically efficient than An27 (in uricase activity from the last common ancestor of mammals are identical and labeled as An32/33. The numbers in brackets between nodes represent the number of amino acid replacements that occur along each gene (Fig. 1). To better understand why the human uricase is not simply replacing the two premature stop codons present in the modern pig uricase (considered to be one of the more active divergences whose descendents ultimately trace to humans.

To test this alternative hypothesis, we resurrected ancient uricase

**Results**

To understand the role ancient uricases would have in fructose metabolism (24), whereas the inhibition of uricase enhances fat stores in part because of its ability to generate uric acid during response to fructose, a major nutrient in fruits that were a primary component of our ancestors’ diet (46). Uricase-expressing cells (Fig. 3 E) showed a significant increase in uric acid accumulation (Fig. 3 F) compared to control cells (Fig. 3 E). The increase in uric acid accumulation correlated with the inhibition of ACC kinase, and its target protein ACC, in response to fructose (Fig. 3 G).

**Fig. S1**

Kratzer et al., 2014, PNAS

**Fig. S2**

Kratzer et al., 2014, PNAS
Uricase evolution in primates

Why was uricase lost in the great apes?

Thrifty genes vs. “drifty” genes

Kratzer et al., 2014, PNAS

Oceanic paleoenvironments

Adaptive protein evolution
Paleoenvironments (EF-Tu)

- Resurrected proteins can provide clues about the temps at which ancient organisms lived
- EF-Tu, an elongation factor crucial for protein synthesis in all cells throughout evolutionary history
- Present day organisms have EF-Tu’s which are highly correlated to temp at which organisms live
- Express resurrected gene into E coli, measure thermostability ($T_m$) of proteins using CD
- Bacterial ancestors appear to be thermophilic (60-80 deg C)
Gaucher et al., 2008 Nature (451):704
Rhodopsin evolution:
Nocturnality of early mammals?

- Reptiles
- Amniota
- Monotrems
- Marsupials
- Theria
- Placental

Previous Hypothesis about early mammals:
1. Living in Nocturnal Niche (Crompton, Taylor and Jaggar 1976 Nature)

Visual cycle: conformational changes in rhodopsin

- 11-cis retinal
- Opisn
- Rhodopsin (500 nm)
- Bathorhodopsin (530 nm)
- Luminorhodopsin (497 nm)
- Meta I (478 nm)
- Meta II (390 nm)
- G-protein Transducin

Light

11-cis retinal

all-trans retinal
Spectroscopic assays of rhodopsin function

In vitro expression & purification

Rhodopsin spectral tuning

Amniote Rhodopsin

$\lambda_{MAX} = 500 \text{ nm}$

Amniote Rhodopsin

$\lambda_{MAX} = 500 \text{ nm}$

Lifetime of activated state

Amniote Rhodopsin

$\tau_{1/2} = 16.5 \text{ min}$

Kinetic rates of light-activated rhodopsin lifetimes

Amniote Rhodopsin

$\tau_{1/2} = 16.5 \text{ min}$

Mammalian Rhodopsin

$\tau_{1/2} = 27.5 \text{ min}$

Theria Rhodopsin

$\tau_{1/2} = 27.5 \text{ min}$

Increased lifetime of activated state of rhodopsin in mammalian and therian ancestors

Bickelmann et al. 2015, Evolution
Synaptic neurotransmission

Glutamate receptors: Excitatory synaptic neurotransmission
Reconstructed ancestral AA-binding GPCR

Ancestral ligand binding site of AA-binding GPCR

Comparison of potency (EC50) Glutamate R agonists

<table>
<thead>
<tr>
<th>Agonists</th>
<th>S.24 receptor</th>
<th>Ancestral receptor</th>
<th>Q78R/31062</th>
<th>Q74R</th>
<th>mGluR1a</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-CCG-1</td>
<td>4.9 ± 1.2</td>
<td>3.5 ± 0.8</td>
<td>1.3 ± 0.1</td>
<td>24.7 ± 1.9</td>
<td></td>
</tr>
<tr>
<td>kainate</td>
<td>183.7 ± 22.5</td>
<td>0.8 ± 0.3</td>
<td>3.2 ± 0.4</td>
<td>6.4 ± 0.3</td>
<td>4.4 ± 0.9</td>
</tr>
<tr>
<td>quisquulate</td>
<td>230.0 ± 18.0</td>
<td>0.8 ± 0.2</td>
<td>3.5 ± 0.8</td>
<td>8.3 ± 0.4</td>
<td>0.4 ± 0.2</td>
</tr>
<tr>
<td>DHPG</td>
<td>23.4 ± 1.3</td>
<td>3.8 ± 1.1</td>
<td>4.3 ± 0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-SOP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Kuang et al., 2006
PNAS (103):14050
Coral pigments

Reconstructed GFP-like proteins from coral

great star coral (Montastraea cavernosa)

Ugalde et al., 2004
Science (305):1433
Conclusions: Ancestral reconstruction

Ancestral reconstruction approaches can offer a window into the past in studying ancient adaptive shifts in protein function.

Computational analyses can be used to generate specific evolutionary hypotheses that can then be tested experimentally.

Experimental approaches should not be viewed as applications of computational methods, instead serve to extend the hypothesis testing framework to study the evolution of protein function.

Need for more interaction between computational and experimental methods in order to provide better insight into both approaches in the study of molecular evolution.