Determinants of the rate of protein sequence evolution

Jianzhi Zhang and Jian-Rong Yang
Published online 09 June 2015
Evolutionary rate (R)

• Measure of dynamics of change in a lineage

• Estimated for a pair of species as number of substitutions per site (d) divided by time of divergence (T)
  • d: Fraction of differing amino acid positions, corrected to account hidden substitutions
  • T: irrelevant (constant) if comparing proteins between given set of species

• Historical findings
  • Molecular clock hypothesis – Zuckerkandl & Pauling 1965
  • Neutral theory of evolution – Kimura 1983
Rate determinants

• Neutral theory

\[ k = \mu \times p \]

- \( k \): protein evolutionary rate
- \( \mu \): rate of mutation
- \( p \): proportion of neutral mutations

• \( p \): determined by functional constraint, which in turn is inferred from \( k \) (Circularity)
• Functional importance, expression level?
Functional importance

• Fitness advantage to the organism
  • More important the protein slower it evolves
  • Measured by fitness reduction upon gene deletion

• Empirical results –
  • Hurst & Smith 1999
    • 175 essential and non-essential mouse genes
    • No significant difference in evolutionary rates
  • Hirsh & Fraser 2001
    • 500 non-essential genes in yeast
    • Weak negative correlation between fitness reduction and evolutionary rates

Is functional importance irrelevant?

• Possible confounders
  • Laboratory vs natural environment
  • No strong correlation in 400 different laboratory conditions

• Predictive power
  • Two random yeast proteins
    • Slower evolving protein 54% more likely to be more functionally important
  • Two yeast proteins rank-separated by 95% of proteins
    • Slower evolving protein 81% more likely to be more functionally important
Expression level

• Pal et al (2001)
  • Strong E-R (Expression-Evolutionary Rate) anticorrelation in yeast
  • True to varying extents across all three domains of life
  • Observed across tissues

• Drummond et al (2005)
  • Stronger signal when mRNA concentrations are used relative to protein concentrations

• Correlation remains after controlling for functional importance

Fig 1.

**Bacteria**
- *Escherichia coli*
  - Log gene evolutionary rate vs. log mRNA expression level
  - \( p = 0.47 \)
  - \( P < 10^{-11} \)

**Archaea**
- *Sulfolobus solfataricus*
  - Log gene evolutionary rate vs. log mRNA expression level
  - \( p = 0.30 \)
  - \( P < 10^{-10} \)

**Plants**
- *Oryza sativa*
  - Log gene evolutionary rate vs. log mRNA expression level
  - \( p = 0.24 \)
  - \( P < 10^{-11} \)

**Fungi**
- *Saccharomyces cerevisiae*
  - Log gene evolutionary rate vs. log mRNA expression level
  - \( p = 0.49 \)
  - \( P < 10^{-5} \)

**Animals**
- *Caenorhabditis elegans*
  - Log gene evolutionary rate vs. log mRNA expression level
  - \( p = 0.20 \)
  - \( P < 10^{-10} \)
Misfolding avoidance hypothesis

• Protein misfolding is cytotoxic and reduces fitness
  • Errors in translation leading to reduced stability

• Higher expressed proteins under stronger pressure to evolve translational robustness
  • Constrains sequence evolution
  • Leads to lower mutation rates and hence E-R anticorrelation

• Other validated predictions for higher expressed proteins
  • Higher folding stability
  • Sites critical for stability more conserved
Misinteraction avoidance hypothesis

• Misfolding induced by surface residues is weaker than core residues
  • But, surface residues show E-R anticorrelation as well

• Surface residues critical for interaction

• Number of misinteracting molecules higher for highly expressed proteins
  • Stronger constraint on sequence evolution
  • Leads to lower mutation rates and hence E-R anticorrelation

• Misfolding and misinteraction avoidance hypotheses – complementary insights
mRNA folding requirement hypothesis

- mRNAs of highly expressed genes have stronger folding
  - Not a product sequence level differences
  - Random mutations more harmful
  - Lower substitution rate (confirmed in yeast)

- Why stronger mRNA folding?
  - Stronger the folding, slower the ribosome elongation
  - Higher translational fidelity
Fig 2.

Misfolding avoidance hypothesis
Selection against errors in translation and folding

Misinteraction avoidance hypothesis
Selection against errors in translation and protein–protein interaction

mRNA folding requirement hypothesis
Selection against errors in translation

- Number of misfolded molecules vs. Expression level
- Strength of selection against misfolding vs. Expression level
- Translational efficiency and folding stability vs. Selection
- E–R anticorrelation

- Number of misinteracting molecules vs. Expression level
- Strength of selection against misinteraction vs. Expression level
- Propensity for misinteraction vs. Selection
- E–R anticorrelation

- Requirement for translational fidelity vs. Expression level
- Required degradation speed vs. Expression level
- mRNA folding strength vs. Expression level
- E–R anticorrelation
Expression cost hypothesis

- **C**: Cost, **B**: Benefit, **ε**: abundance/expression

- Optimal abundance: \( B'(\varepsilon) = C'(\varepsilon) \)

- Mutation that decreases activity by a fraction \( q \) \( \Rightarrow \) loss of \( q \varepsilon \) molecules

Fig 3.
Expression cost hypothesis

• Lacking empirical evidence
  • How to quantify expression cost?
  • How does cost scale with protein length?
  • What if cost per molecule is not constant for proteins?
<table>
<thead>
<tr>
<th>Correlates</th>
<th>Properties of faster evolving proteins</th>
<th>Organisms</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene expression level</td>
<td>Lower expressions</td>
<td>Bacteria, archaea and eukaryotes</td>
<td>9,14,34,37</td>
</tr>
<tr>
<td>Functional importance</td>
<td>Lower importance and higher dispensability</td>
<td>Yeast and mammals</td>
<td>27,30–32</td>
</tr>
<tr>
<td>Expression breadth among tissues</td>
<td>Lower expression breadth and higher tissue specificity</td>
<td>Mammals</td>
<td>32,100</td>
</tr>
<tr>
<td>Expression timing in development</td>
<td>Expression in late embryogenesis and adulthood</td>
<td>Zebrafish</td>
<td>101</td>
</tr>
<tr>
<td>Promoter and gene body methylation</td>
<td>Higher levels of promoter methylation but lower gene body methylation levels</td>
<td>Mammals</td>
<td>102</td>
</tr>
<tr>
<td>Chaperone targeting</td>
<td>Higher levels of chaperone targeting</td>
<td>Bacteria and eukaryotes</td>
<td>103,104</td>
</tr>
<tr>
<td>Protein subcellular localization</td>
<td>Higher tendency to be extracellular</td>
<td>Yeast and mammals</td>
<td>105</td>
</tr>
<tr>
<td>Codon usage bias</td>
<td>Weaker codon usage bias</td>
<td>Bacteria and eukaryotes</td>
<td>14</td>
</tr>
<tr>
<td>Distance from the origin of replication</td>
<td>Larger distance</td>
<td>Bacteria and archaea</td>
<td>106,107,108</td>
</tr>
<tr>
<td><strong>Pleiotropy</strong></td>
<td><strong>Lower pleiotropy</strong></td>
<td><strong>Eukaryotes</strong></td>
<td><strong>57</strong></td>
</tr>
<tr>
<td>Protein–protein interaction network properties</td>
<td>Lower connectivity, closeness and betweenness</td>
<td><strong>Eukaryotes</strong></td>
<td>58,109,110</td>
</tr>
<tr>
<td>Metabolic network property</td>
<td>Lower flux and connectivity</td>
<td>Yeast</td>
<td>111</td>
</tr>
<tr>
<td>Regulatory network properties</td>
<td>Higher centrality</td>
<td>Yeast</td>
<td>112</td>
</tr>
<tr>
<td><strong>Targeting by microRNAs</strong></td>
<td>Fewer types of targeting microRNA</td>
<td>Mammals</td>
<td>55</td>
</tr>
<tr>
<td>Gene compactness</td>
<td>Shorter introns and untranslated regions</td>
<td>Mammals</td>
<td>32</td>
</tr>
<tr>
<td>Protein length</td>
<td>Longer proteins</td>
<td>Yeast and mammals</td>
<td>113,114</td>
</tr>
<tr>
<td>mRNA folding</td>
<td>Weaker mRNA folding</td>
<td>Bacteria and eukaryotes</td>
<td>12</td>
</tr>
<tr>
<td>GC content</td>
<td>Lower GC content</td>
<td>Mammals</td>
<td>113</td>
</tr>
<tr>
<td><strong>Domain structure</strong></td>
<td><strong>Lower density of domains</strong></td>
<td><strong>Animals and plants</strong></td>
<td><strong>54</strong></td>
</tr>
<tr>
<td>Protein disordered regions</td>
<td>More-disordered regions</td>
<td>Bacteria and eukaryotes</td>
<td>115</td>
</tr>
<tr>
<td>Protein structural designability</td>
<td>Higher inter-residue contact density and higher fraction of buried sites</td>
<td>Yeast</td>
<td>114</td>
</tr>
<tr>
<td>Protein conformational diversity</td>
<td>Lower conformational diversity</td>
<td>Mammals</td>
<td>116</td>
</tr>
</tbody>
</table>
Broader implications

• Functional constraint
  • Functional importance only a minor determinant
  • Creation of toxicity more important

• Misfolding/misinteraction in genetic diseases

• Integrative approach to study multiple factors constraining evolutionary rates