Repetitive Patterns in Proteins

Tutorial
Repetitive patterns in proteins

- The reuse of *domains* in various proteins
- The recurrence of *super-SSE* in different domains
Protein domains and super-secondary structure elements

- Protein domains
  - An autonomously folding polypeptide
  - Modular re-use in different contexts
  - Classification schemes are based on domains not entire proteins

- Domains themselves are formed by a set of recurring super-secondary structure elements

[Soding & Lupas, Bioessays 2003]
Examples of protein structural repeats

[Andrade et al. JSB 2001]
Why is this stuff helpful for You?

• About 25% of all eukaryotic proteins contain repeats
• Internal repeats often correspond to structural or functional units in proteins
Duplication is central to the diversification of proteins

• The principle of modularity (duplication/amplification, recombination) operates at all levels of biological organization:
  – Full genomes
  – Entire operons
  – Single Genes
  – (Domains and sub-domain sized fragments)
• Effective path to
  – increased complexity and
  – more adapted proteins,
    • the duplicated copy is free to evolve a novel function

en.wikipedia.org/wiki/File:Gene-duplication
Duplication accompanied by fusion generates novel proteins

• Genetic Mechanisms:
  – Replication slippage
  – Illegitimate recombination
  – Crossover during sexual recombination (“exon shuffling”)
  – (Retro)-Transposition (replicative recombination)

• These processes result in
  – novel domain compositions
  – circularly permuted proteins (includes loss), or
  – **repetitive proteins**
Internal structural symmetry within domains suggests an origin through successive duplications

- Individual repeats span the range from near sequence identity to complete dissimilarity, where the origin through repetition can only be seen on the structural level.
Repetitive patterns are found in different classes of proteins

- Fibrous proteins
- Solenoid proteins
- Membrane proteins
- Globular proteins

- An evolutionary path from “simple” scaffold proteins to fully differentiated enzymes
- Evolution of rather complex and well adapted molecules from smaller units
Fibrous proteins are generated by the repetition of short peptide segments

\[
\begin{align*}
\text{abcdefg} \\
\text{LEEIVNQ} \\
\text{LNIYQSQ} \\
\text{VELIQQQ} \\
\text{MEAVRAT} \\
\text{ISELEIL} \\
\text{EKTLSDI} \\
\text{MESIKSQ} \\
\text{KNELEST} \\
\text{LQKMG} \text{EN} \\
\text{LRKITDI} \\
\text{MMKLSPQ} \\
\text{AEELLKK}
\end{align*}
\]
Monotonous repetition of one Super-SSE gives rise to open-ended solenoid proteins

REYITSLDLSANELRDIDALDLSQKCCISVH
LEHLEKOLEHQNALTSPFPQQLCET
LKSLTHLDHLSNKFTSFPSYLLK
MSCIANLDVSRNDIGPSVVLDPPTVK
CPTLKQFNLSYNQLSFVPENLTVD
VEKLEQLILEGNKISGICSPLR
LKEKILNLSKHNISSSLSENFLEA
CPKVESFSARMNFLAAMPFL
PPSMTILKLSQNKFSCIPEAILN
LPHLRLSDMSSNDIQYLPGPAHWK
SLNLRELLFSHNQISILDLSEKAYL
WSRVEKLHLSHNKLEIPPEIGC
LENLTLSDLVSYNLELRSFPNEMGK
LISKIWDLPLDELHHLNFDFK
Membrane proteins contain repetitive patterns
Some frequently occurring protein folds have internal symmetry

- β- trefoil
- Jelly-roll
- Immunoglobulin fold
- TIM-barrel
- Ferredoxin fold
- Updown bundle
Classification of Repeats

Based on the length of the repeated unit:
- Short:
  - One amino acids (e.g. polyQ in Huntingtin)
  - Three aa (Cold shock protein)
  - Seven aa (Coiled coils)
- Intermediate
  - Super-secondary structure elements
    - \((\alpha\alpha/\beta\beta\)-hairpins, \(\beta\alpha\beta\)-elements, ca. 30-40 aa\)
- Long
  - Entire Domains (ca. 100-200 aa)

Based on theoretical limit of repeat number
- Open
- Closed
Detection of Repeats

• Challenging
  – Sequence degradation
  – Non-integer multiples
  – Non-coincident boundaries
  – Tandem vs. non-tandem

• Pattern is known
  – Identification of specific patterns in fibrous proteins (Coils)
  – Search against databases of known repeats
    • REP: http://www.bork.embl.de/~andrade/papers/rep/search.html
      (Ankyrin, Armadillo, HAT, HEAT, HEAT_AAA, HEAT_ADB, HEAT_IMB, Kelch, Leucin Rich Repeats, PFTA, PFTB, RCC1, TPR, WD40)
    • Mocca (Multiple OCCurrences Analysis)
    • TPRpred @ toolkit

• Pattern is unknown
  – De novo detection of internal sequence repeats by comparing the protein sequence to itself (HHrepID)
The secret of the evolutionary success of repetitive proteins

- Problems: Only very few polypeptide sequences are capable of folding; protein folds are not very stable
- Repetition intrinsically promotes stability through the periodic recurrence of favorable interactions
- Modular reuse of already established components allows for a stepwise increase in complexity (emergence)

**Table 1.** Superfolds and the fraction of their residues contained in the supersecondary structure elements $\alpha\alpha$, $\beta\beta$, $\beta\alpha\beta^{(21)}$

<table>
<thead>
<tr>
<th>Fold</th>
<th>Sequence*</th>
<th>Structure</th>
<th>Number of superfamilies (%)</th>
<th>% Supersecondary structure content</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-trefoil</td>
<td>+</td>
<td>+</td>
<td>2 (0.1)</td>
<td>83</td>
</tr>
<tr>
<td>Jelly roll</td>
<td>-</td>
<td>+</td>
<td>17 (1.2)</td>
<td>47</td>
</tr>
<tr>
<td>Immunoglobulin-like</td>
<td>-</td>
<td>+</td>
<td>55 (4.0)</td>
<td>67</td>
</tr>
<tr>
<td>TIM-barrel</td>
<td>+</td>
<td>+</td>
<td>28 (2.0)</td>
<td>82</td>
</tr>
<tr>
<td>Ferredoxin-like</td>
<td>+</td>
<td>+</td>
<td>65 (4.7)</td>
<td>38</td>
</tr>
<tr>
<td>Updown bundle</td>
<td>+</td>
<td>+</td>
<td>17 (1.2)</td>
<td>90</td>
</tr>
<tr>
<td>OB fold</td>
<td>-</td>
<td>-</td>
<td>16 (1.1)</td>
<td>77</td>
</tr>
<tr>
<td>UB-roll</td>
<td>-</td>
<td>-</td>
<td>16 (1.1)</td>
<td>55</td>
</tr>
<tr>
<td>Globin-like</td>
<td>-</td>
<td>-</td>
<td>4 (0.3)</td>
<td>88</td>
</tr>
<tr>
<td>Doubly wound</td>
<td>-</td>
<td>-</td>
<td>122 (8.8)</td>
<td>68</td>
</tr>
<tr>
<td>All superfolds</td>
<td>-</td>
<td>-</td>
<td>342 (24.7)</td>
<td>65</td>
</tr>
<tr>
<td>All folds</td>
<td>-</td>
<td>-</td>
<td>1386 (100)</td>
<td>62</td>
</tr>
</tbody>
</table>
Why is this stuff helpful for You?

• About 25% of all eukaryotic proteins contain repeats
• Internal repeats often correspond to structural or functional units in proteins
  – Therefore, methods capable of identifying diverged repeated segments or domains at the sequence level can assist in:
    • predicting domain compositions
    • predicting domain boundaries
    • inferring hypotheses about function and mechanism
    • investigating the evolution of the protein of interest
• Further more:
  – Design constructs in a smarter way and do more successful experiments
  – Get a deeper understanding of how the complexity of “modern” proteins evolved