Sequence Motifs: Highly Predictive Features for Protein Function Prediction

Asa Ben-Hur and Douglas Brutlag
Department of Biochemistry, Stanford
Background

- Proteins participate in most of the biochemical processes in the cell
- SwissProt: Protein sequence database. Contains ~140K sequences
- Enzymes: facilitate chemical reactions
- Enzyme Commission (EC) numbers: n1.n2.n3.n4
- SwissProt contains 35K enzymes which belong to ~750 EC classes
Similarity / Representation

- **Similarity:**
  - Weighted edit distance: Smith-Waterman and BLAST methods
  - Model-based, e.g. HMM (Haussler et al.)
  - Fisher kernels (Jaakkola et al.)

- **Vector-space representation:**
  - Extract a set of properties (amino acid counts etc.)
  - Represent a sequence in the space of all $20^k$ k-mers (spectrum and mismatch kernels, Leslie et al.)
  - **Motif composition**
Protein Sequence Motifs

- Evolutionarily conserved sequence elements
- Represented as regular expressions or as position-specific scoring matrices
- Known to be part of protein functional sites:
  - Catalytic sites
  - Binding sites

Syntax:

```
k[ilmv]...hq
```

Snippet of a Multiple sequence alignment:
```
VIGCANCHDDKT-
TIGCYNCHDDKS-
NLGCADCHNTASP-
---ECDSCHTPD--
---GCDSCHVSDK-
---SCQSCHAKP--
---GCESCHKDG--
---QCLNCHSPEN-
EVGCIDCHVDVN-
---DCVGCCHVDGF
---DSVLCHISV--
```

Motifs:
```
C  .  C  H  
GC  .  C  H  .  D
```
Computing Motif Composition

Represent motif database in a **TRIE** with motifs in leaf nodes
The Motif Representation

- A “bag of motifs” representation of a protein sequence:

\[ \Phi(x) = (\phi_m(x))_{m \in M} \]

- A high dimensional feature vector: motif database can contain several hundred thousand motifs

\[ K(x, x') = \Phi(x) \cdot \Phi(x') \]

The motif kernel is a linear kernel that essentially counts the number of motifs two sequences have in common.
Assessing Motifs as Features

For each class of enzymes we compute a statistic for each feature:

\[ \max_i P(m_i \mid \text{class}) - P(m_i \mid \text{out of class}) \quad \max_i \text{Specificity}(m_i) \]
Feature Selection Results

Feature selection using the $L_0$ (multiplicative update) method of Weston et al. compared with SVM trained on all features:

# features for each class  Balanced Success Rate:
Classification Results

- KNN works very well:
  - Success rate on all data: 0.94 (same as SVM)
  - One-against-rest comparison with SVM:

Area under ROC50 curve  Balanced Success Rate
**Conclusion**

- Motifs: highly discriminative features for predicting the function of a protein
- Can provide low dimensional, interpretable classifiers
- Domain knowledge required

Things I haven’t mentioned:
- Discrete motifs vs. scoring matrices
- Custom motif databases for enzyme classification