A Sensitive Profile-Profile Comparison Tool based on Information Theory - Yona et al

Arvind Gopu

agopu [at] cs[dot]indiana[dot]edu

November 13th 2003

Bioinformatics Research Group

School of Informatics

Indiana University, Bloomington IN

In a Nutshell...

A profile-profile comparison tool which:

- Detects weak similarities between protein families - even in the twilight zone of similarity
- Allows for gaps
- Produces alignments that are in agreement with structural alignments
- Is much more sensitive than existing tools like IMPALA, PSI-BLAST

Existing Methods - Background

- Twilight Zone: 20-30% Sequence identity; Pairwise and existing profile vs seq tools fail to find similarities in this zone
- But with even higher mutation function and fold might be retained. How to capture these?
- Even iterative searching (PSI-BLAST) will not help
  - Strict parameters will result in missing divergent seq relationships
  - Lenient parameters will cause unrelated seqs to sneak into the results
- One main reason why both problems occur is – similarity and statistical significance are not considered together

Existing Methods - Background ...

Existing profile tools:

- Include PSI-BLAST, HMM based tools, CASP, SAM T*, etc.
- Most of these tools are already quite good (especially PSI-BLAST – used extensively) but yet they miss remotely related seqs
- Some of them incorporate structure information in an effort to overcome this - partially successful.
- Common feature? Seq compared against a model

Proposed Algorithm

- Does a model-model (or rather profile-profile) comparison - claimed (and shown) to be more effective in detecting remotely related seqs
- Model Info:
  - Uses dynamic programming
  - Information theory based measure for similarity of prob. distrs.
    - Model Independent
    - Combines similarity score with statistical significance to come up with ultimate score
    - Local sequence similarity also taken care off.
- Most similar work is the CASP project

Proposed Algorithm ...

- What is a profile anyway?
  - Representation of group of related protein seqs based (usually) on MSAs
  - Once MSAs are generated, count amino acids at each column-wise and transform counts to probabilities .
  - Prob. distr. denote likelihood of observing a particular amino acid at a specified position
  - Pseudocounts are usually added - Laplace rule
- Above mentioned prob. distrs. can be thought off as a matrix of 20 rows and n columns where n corresponds to the number of columns in the sequences

Proposed Algorithm ...

Data sets: Several classifications of protein sequences available - SCOP, CATH, FSSP/DALI, etc. Yona et al have used the SCOP database in their analysis

- Manually curated
- Breaks proteins into domains – eliminating multidomain proteins
- SCOP 1.50 had 23,780 protein domains classified into 1287 families, 814 super families, 545 folds and 7 classes

Proposed Algorithm ...

- Choose seed - sequence whose average distance from all other members of the family is smallest.
- Use PSI-BLAST with this seed against all other seqs in family
  - If there is just one seq or if PSI-BLAST fails to produce a profile, represent it by profile generated from BLOSUM62
  - No unrelated seqs since it's intra family BLAST searches; no danger of false positives being included in profile and less sensitive to tweaking of parameters
- For profile-profile comparison use dynamic programming algorithm and assign profile-profile alignment score – explained in the following slides

Proposed Algorithm ...

Divergence Score:

For two profiles $P = p_1 p_2 \ldots p_n$ and $Q = q_1 q_2 \ldots q_m$

where m and n are the lengths of the profiles ie number of columns and

p_i and q_j are probability distributions over the 20 letter alphabet of amino acids...

The Kullback Leibler (KL) divergence is:

(1) $$D^{KL}[p_i||q_j] = \sum_k p_{ik} \lg \frac{p_{ik}}{q_{jk}}$$

Proposed Algorithm ...

Divergence Score:

- Above defined KL divergence not good enough - Asymmetric and unbounded
- Better measure?

  Yona et al have use Jensen-Shannon divergence for probability distributions - Symmetric as well as bound (between 0 and 1)

Proposed Algorithm ...

For two profiles p and q, and for every $(0 \leq \lambda \leq 1)$

The Jensen-Shannon (JS) divergence is defined as:

(2) $$D_{\lambda}^{JS}[p||q] = \lambda D^{KL}[p||r] + (1-\lambda)D^{KL}[p||r]$$

where

(3) $$r = \lambda p + (1-\lambda)q$$

is most likely common source distribution of both p and q. Without *A priori* information λ will obviously have to be $1/2$.

Proposed Algorithm ...

Significance Score:

- Match of two prob. distrs. that both resemble a unique distr. should be more significant than two that resemble the overall distr. of amino acids (ie. random)
- So use large sequence databases like SWISSPROT + TrEMBL to define an overall base amino acid distribution (P_0) and define significance score to be the JS divergence of the common source distr. (r) to the base distr (as shown below).

(4) $$S = D^{JS}[r||P_0]$$

Proposed Algorithm ...

Final Score:

Define a final score combining the two above mentioned scores.

$$\text{(5)} \quad Score(p,q) = \frac{1}{2}(1-D)(1+S)$$
$$= \frac{1}{2}(1-D^{JS}[p||q])(1+D^{JS}[r||P_0])$$

Equation (??) shown above has a couple of intuitive properties. This is explained in the following slide...

Proposed Algorithm ...

- For similar distrs. $D \rightarrow 0$; if the common source is far from the background distr (P_0)
  - then $S \rightarrow 1$
  - else $S \rightarrow 0$
- $Score(p, q)$ also is capable of differentiating between:
  - two similar distrs. ($D \rightarrow 0$) that are also individually similar to P_0 ($S \rightarrow 1$) => $Score(p, q) = 1/2$
  - two dissimilar distrs. ($D \rightarrow 1$) that are individually similar to P_0 ($S \rightarrow 1$) => $Score(p, q) = 0$
- Look at results on paper (PDF)

Proposed Algorithm ...

Shift Transformation of score (For local similarity):

- The final score as mentioned falls between 0 and 1
- But local alignment similarity functions need to satisfy:
  - $mean(score(p, q)) < 0$
  - $max(score(p, q)) > 0$
- They came up with an empirical value to shift lying between 0.42 and 0.5; They have analysis of their algorithm's performance for various such values

Optimization of Parameters

- Used test set of 120 families - each of which had at least two related families within the same SCOP super family
- Given a parameter set, compute profile-profile similarities with all 1287 families, sort results and count # of true positives before first false positive (definition of FP important here)
- Aim for not only max sensitivity but also highest accuracy (obviously!)
- Compare profile alignments with structure of family seed generated using structure tools - Structal and CE

Optimization of Parameters ...

- Consider only alignments which have E-value ($e \leq 1$)
- Set of indices defined:

$$N_{aligned}, Q_{shift}, N_{agreement}, Q_{modeller}, Q_{developer}$$

- Look at figure 4 in the paper (page 1263) for more details

Statistical Significance

- Two baseline empirical distrs. established
  - One based on matches of profiles of totally unrelated families (belong to different SCOP classes) – Used to assess significance of match for ANY two given profiles (Figure 5a on paper)
  - Second based on matches of profiles of a specific family - to assess significance of matches with the particular profile (Figure 5b on paper)
- Extreme value distribution fit to both of the above distrs.

Performance Evaluation

- Please look at tables shown Pages 1266 and 1267 for comparison of results as well as explanation of the same.
- Especially the part where they explain how their method works around the problem seq to model comparison tools suffer from (Page 1266, right column - middle)

Conclusion

- Results shown in the paper do indicate that the proposed method is much better in detecting remotely related families

References

To be filled in – later!