Alignment-Free Classification and Comparison of Biological Sequences and Structures

Raffaele Giancarlo

Dipartimento di Matematica, Università di Palermo

Outline

- Comparison and Classification of Sequences and Structures
- Alignment Methods (Main Ingredients)
- Alignment-Free Methods
  - Theory and Practice
  - Evaluation Methodology
  - Conclusions Based on Experiments
  - Software

Bibliography

- Huge and growing…
- I have some of it on-line

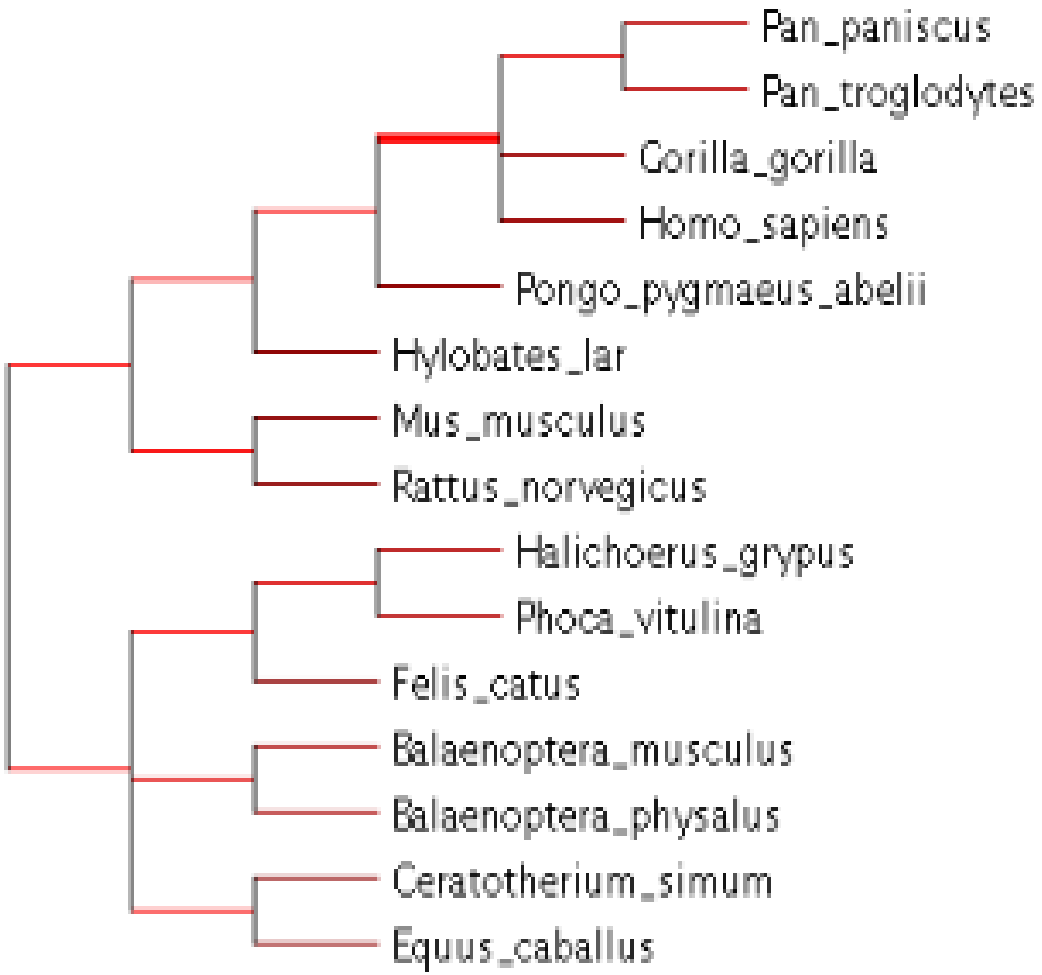
Comparison and Classification-General

- Inference of homology and function
  - **Basic Axiom of Computational Biology: Guilt by Association** *A high similarity among objects, as measured by mathematical functions, is strong indication of functional relatedness and/or common ancestry…**Not always***
- Basic Problems
  - Definition of good similarity/distance functions
  - Development of efficient algorithms for their computation

BOTH DIFFICULT PROBLEMS

Comparison and Classification-General

- Example

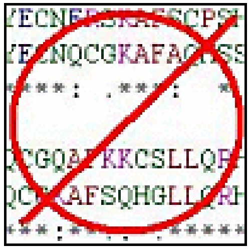


Comparison and Classification-General

- Data that can be represented as strings
  - This talk has some relevance
- More Complex Data
  - Protein Structures
    - The Ten Most Wanted Solution in Bioinformatics [Tramontano]
  - Networks
    - Sharan and Ideker

Comparison and Classification-General

- Basic Ingredient: Similarity/Distance functions between strings
  - Two Approaches:
    - Functions based on Alignment Methods
    - Functions not based on Alignment Methods



Alignment-Free Methods

Alignment Methods- Basics

- Two Strings
  - Global alignments

    a x a b – c s
    a x - b a c s

  - Local alignments

    X= pqraxabcsstvq ; Y= xyaxbacsll

  - In both cases, one gets a similarity value stating how similar two strings, or parts of them, are.

Alignment Methods- Basic Algorithms

- Dynamic Programming: NW and SW
- Heuristics: FASTA, BLAST, PATTERHUNTER…
- All Algorithms need a scoring scheme
    - Proteins:
        - PAM, BLOSUM substitution matrices, ad hoc gap penalties
    - DNA:
        - Heuristic schemes

Alignment Methods-Limitations

- No shuffling or interchange operation allowed
  - They do not account for recombination with shuffling
- Their performance does not scale well with Data Set size
  - Difficult to use on a genome-wide scale
- Sensitivity depends on choice of weight matrices
  - Difficult to use in the “twilight zone” : sequence identity <20%

Alignment-Free Methods

- Similarity of two strings is assessed based only on the DICTIONARY of substrings that apper in the strings, irrespective of their relative position
  - Lipari, abracadabra, ababraracad
- Advantages:
  - No parameter setting, no training, no learning
    - Towards Parameter-Free Data Mining, Lonardi et al.
  - Speed and Scalability
    - Time linear in the size of the input

Alignment-Free Methods

- Computational Approaches:
    - Explicit Collection and Use of Word Statistics, either exact or approximate
        - Similarity/distance of two strings reduces to similarity/distance of points in high-dimensional geometric spaces
    - Implicit Collection and Use of Word Statistics
        - Kolmogorov Complexity, Information Theory and Compression

Explicit Collection of Word Statistics

- See paper by Vinga and Almeyda
- Related Issues
  - Kernel Functions in SVM- Protein Classification
  - Linguistic Complexity- Coding/NonCoding Regions
  - Compositional Complexity-Coding/Noncoding Regions
- See papers by Bolshoy and Konopka

Implicit Collection of Word Statistics

- Intuition: Similarity is captured by quantifying “how easy” it is to describe x, given y
- Example: abraabraabra | abra
  - Kolmogorov Complexity and/or Data Compression
- Similarity via Relative Compressibility

Universal Similarity metric (USM)

$$USM(x,y) = \frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}}$$

- Universality here is a very powerful concept: USM is a lower bound, and therefore a good estimator, of any computable distance/similarity function

- Problem:
  - USM(x,y) is based on Kolmogorov Complexity that is non- computable in the Turing sense.

Universal Similarity Metric

- Resort to compression

- Given compression algorithm C, $K(x)$ can be approximated by $|C(x)|$, $K(x,y)$ by $|C(xy)|$ and $K(x|y^*)$ by $|C(xy) – C(x)|$.

- In practice, USM become a methodology that depends critically on the choice of compression algorithm.

Approximations of USM

- Given compression algorithm, three general formulas to approximate USM

$$UCD(x,y)=\frac{\max\{C(xy)-C(x),C(yx)-C(y)\}}{\max\{C(x),C(y)\}}$$

where

$$NCD(x,y)=\min\{NCD_1(x,y),NCD_1(y,x)\}$$

$$NCD_1(z,w)=\frac{C(zw)-\min\{C(z),C(w)\}}{\max\{C(z),C(w)\}}$$

$$CD(x,y)=\frac{\min\{C(xy),C(yx),C(x)+C(y)\}}{C(x)+C(y)}$$

Lempel-Ziv Complexity

- Complexity of a finite sequence, given knowledge of another:, via LZ77 Parsing:
- abra abra,abra, abra, l,i,p,ar,i
- Avarage Common Substring:--Ulitsky et al.

Experiments: General Conclusions

- Vinga et al +Ferragina et al.+ Ulitsky et al
  - Alignment free methods are good filtering techniques for classification and assessment of similarity
  - They are efficient and scale well with data set size
  - They can be successfully applied also to protein structures, not only when the the domain of interest is
    in string format
  - Reliable philogeny reconstruction on a genomic and proteomic scale
  - The “memory” of a compressor is important for genomic data, much less so for protein representations

Software

- See papers by Vinga et at.
  - url: http://bioinformatics.musc.edu/resources.html
- See peper by Ferragina et al.
  - url: http://www.math.unipa.it/~raffaele/kolmogorov/
- ProCKSI- Barthel et al.

…And More to Come – Part I

- Biological Network Comparison
  - Based on Alignments –Sharan and Ideker
  - (The first) Alignment-Free Method- Chor and Tuller
    - Minimum Description Length Principle

…And More to Come – Part II

- The Quest for a mathematical definition of “Biological Information”.
    - State of the Art: P. Godfrey-Smith and K. Sterelny
    - Latest: Galas et al (2008).– Set Based Complexity and Biological Information
        - Kolmogorov complexity and Data Compression strike again!!!

…And More To Come – Part III

- R. Giancarlo, D. Scaturro and F. Utro, Textual Data Compression and The –omic Sciences: A Synopsis,

  Manuscript prepared for Biojnformatics, ready for submission

Good News: Compression is pervasive

Bad News: Its use and tools coming from it is totally disorganized- very low impact