Cognate Recognition, Phylogeny and Alignment

7.11.15

Background and Motivation: Stochastic models of sequence evolution are central in modern genetics and have over the last decades included not only change of single characters [substitutions] but also insertion-deletions (indels). These combined substitution-indel models allow a statistical approach to alignment and phylogeny inference [Phylogenetic Statistical Alignment].

Substution models have over the last 4 decades undergone a significan evolution towards more complexity and realism, by including dependence, position heterogeity and annotion [hidden states]. Indel models has yet to undergo this development. Although it is clear what is needed, there are so many computational challenges in large scale application of the simple models [TKF91 TKF92] that this has attracted all the effort. But eventually more realistic models will be explored.

Phylogenetic Statistical Alignment has already proven its worth in biology, but there is an obvious application that would be nice to explore – word evolution in linguistics!! In this initial project we will focus on alignment. Statistical alignment should hopefully have the advantage to have several alignments with probability assignments when there is ambiguity, where as early earlier methods would be forced to make one educated guess.

Sometimes word analysis is reduced to statements of being cognate which corresponds to homology in biology. HOUSE [English] and HAUS [German] are cognate, while HOUSE [English] and MAISON [French] are not. Assignment of cognancy is often done manually by a linguistic although tests have been designed [GERHARD something here]. Again there are strong similarities to homology testing for sequences [Hein et al., 2000; Altschul et al., 1990]

There are two large domins of Statistical Alignment in Historical Linguistics – Phylogenetic Alignment & Cognate recognition:

A. Phylogenetic Statistical Alignment: Word evolution is different from DNA evolution and will need some model development. More realistic models are typically computationally more demanding, but the good news here might be that words are much shorter [4-12 characters] while DNA is much longer [10^2-10^8]. Word evolution needs models with these properties:

1. **Correlation between changes in different positions** – can clearly happen if one sound is changed in general at different places in the word.

2. **Swaps** – is not a typical problem in biological sequene analysis, but has been considered in computer scientist [Lowrance and Wagner, 1975].

3. **Positional Heterogeneity** – is a standard problem in sequence analysis and is typically solved by assigning random independent rates for different positions or by HMM/SCFG [Stochastic Context Free Grammars] giving a distribution of correlations on hidden states. The latter creates certain problems if coupled to an insertion-deletion process since this destroys neighbor relationships.

In summary these additional constraints are making the formulation of a dynamic programming solution hard and it must at least be complemented by a "summing over all paths" approach which has been applied successfully in many contexts [Miklos, Lunter & Holmes, 2004; Miklos, 2003].

B. Cognate Recognition (Homology Testing: This problems is very similar to the analogous problem in biological sequence analysis. In testing for this it is natural to use P(w1,w2)/[P(w1)P(w2)] – i.e. how probable is the data (two words) given they are related by a small tree relative to how probable is the data if the two words were independent. The quantitity P(w) is the probability of the word according to the equilibrium distribution in the "word process". The equilibrium distribution of ordinary statistical alignment is the geometric distribution which is bad description of the lengths of words. For instance the most probable word length is zero…. For tree reconstruction purposes the bad equilibrium distribution is of little consequence but for word homology it will be of little consequence. But for word homology testing we must also ask for

4. **Realistic Equilibrium Distribution** – I am sure that an appropriate negative binomial distribution would do. It would be possible to impose such a distribution on the equilibrium distribution. However, changing the whole process to give the right equilibrium distribution would be harder since now each event would have to be biased toward the equilibrium distribution.

As in biology it might actually be an advantage to take two sets of words that each are known to be cognates [**w1** and **w2**] and test if the two subtrees can be merged using P(**w1**,**w2**)/[P(**w1**)P(**w2**)]. The reason that this could be efficient is that aligning a set of words will give information on the evolutionary process of evolution and which positions are conserved in the word (if this was part of the model) and thus would give more information about what that word is. This should lead to a better test.

Mini-Project (10 weeks for a single student)

A mini-project should stay very limited and only explore how existing statistical alignment tools modified to a 41 letter phonetic alphabet behaves. The analogues of PAM matrix (and time back to common ancestor - t), equilibrium frequencies of letters are available from which a rate matrix Q for letter evolution can be derived. To test the possibility of statistical alignment in linguistics it is natural to start out with some very simple data and models sets of increasing difficulty. It will also be be natural to not ask for all extension to statistical alignment listed above, but at first use it in its present form but with an other alphabet [phonetic] and mutation matrix [see figure] obtained from earlier word comparisons.

The program StatAlign [Novak et al.] should be able to do this after suitable modification. Since the test data sets will be small, it might not be realistic to infer parameters for branch lengths, substution/insertion-deletion rates so it would be fully acceptable to set these to qualified guesses.

Alignment Testing:
Data Set 0: A series of obvious cognate pairs here examplified by the word evening in 4 Germanic languages. Danish: afd3n, German: abEnt, Dutch: avont and Swedish: aftun.
Data Set 1: Two set of words from two very similar languages.
Data Set 2: Two set of words from two different languages from the same language family.
Data Sets 3-5: Same as above but more than 2 languages.

Alignment Testing:
The key quantity to calculate for a word pair is R=P(w1,w2)/[P(w1)P(w2)] where a time parameter t must be set analogously to PAM in sequence analysis. Again a qualified choice must be made. Then word pairs must be simulated in P() and the R plotted for the real word pair and for the simulated distribution. If it is in the 1% fractile in the simulated distribution, then one would declare the two words for cognate. Using TKF91 one is naïve but should be done for a starting test.

Extenstions to DPhil

If the first two thirds of the mini-project is succesfull then we will talk with Jaeger, Wahle and Wichman and make a plan for a full Dphil. There is plenty to do and data enough so it is a question of being realistic in how do extend statistical alignment. This would most likely have to be re-implemented since in realistic most simplifying assumptions of statistical alignment are wrong.

Refererences

Swadesh, Morris (1952). Lexico-statistic dating of prehistoric ethnic contacts. Proceedings of the American Philosophical Society 96, 452-463
Hein,J., C.Wiuf, B.Knudsen, Møller, M., and G.Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. (J. Mol. Biol. 302.265-279)
J.Hein, J.Jensen and C.Storm (2003) "Algorithms for Multiple Statistical Alignment" (*PNAS* 100(25):14960-14965.)
Satija, R.,Pachter, L. & Hein, J. (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics 24,1236-1242
Miklós, I., Lunter, G.A. & Holmes, I. (2004) A 'long indel' model for evolutionary sequence alignment. *Mol. Biol. Evol.* 21(3):529–540.
TKF92 Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution.J Mol Evol. 1992 Jan;34(1):3-16.
TKF91 Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences.J Mol Evol. 1991 Aug;33(2):114-24
I Miklós (2003) MCMC genome rearrangement Bioinformatics 19 (suppl 2), ii130-ii137
R. Lowrance and R. A. Wagner. An Extension of the String-to-String Correction Problem. Journal of the Association for Computing Machinery, 22:177–183, 1975.
Altschul SF1, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.
Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

Sources of Words: A series of databases have been collected. http://asjp.clld.org and http://lingulist.de/lingpy/

Comments and Questions: Testing of homology in sequence analysis is for instance done in BLAST and is in general done under pressure for computational efficiency thus the tests are quite simple. PSI-BLAST must have taken some simples steps to allow position weighting in a homology test. I don't know if this ever has been taken any further.


Robin Ryder

Oxford Work This means supervised or intiated by Jotun Hein. A lot of interesting work is going in Oxford that I will not mention here. Might be in future editions of enlarged project description.


Thomas Mailund

I find linguistics and historical linguistics very interesting and have always read books on linguistics regularly. I did a postdoc in 1989-90 on sequence alignment with David Sankoff in Montreal, who was one of the founders of the use of models in mathematical models in historical linguistics.

I have supervised two projects on this topic and I wish there had been opportunity to do more. Both reports are available from the project page.


David Sankoff

The first project was a DTC project in 2006 where I with Thomas Mailund [yes THE THOMAS MAILUND!!!] supervised Robin Ryder who did his PhD in historical linguistics under the supervision of Geoff Nicholls and got the Corcoran Memorial Prize for his PhD report. Robin and I went to Santa Barbara together to attend a Workshop on Historical Linguistics organised by Bernhard Comrie.


Stephen Clark

Together with computational linguist Stephen Clark from Comlab in Oxford [Now moved to Cambridge] we proposed a project on changing a grammar of one language into the grammar of another language. The idea was simple: Get the set of grammatical rules for the two languages and start substituting the rules of the first language with rules of the second language and clearly the "current language" would move from the first language toward the second language. So far so good. We advertised our project and Markus Gerstel chose to do it with us. He was very good, organised and very compentent. I was familiar with grammars for RNA secondary structure which is 3-5 rules and changing them was the topic for another OSSCB project which led to a publication. However, doing it for languages were very different. We found text corpus grammatical rules for English and German. These corpii and their grammars are clearly convention dependent leading to the problem that the same language could be described by different grammars, so the distance of a language to itself could be greater than zero. Then we you delete some rules other rules might be useless or the same sentences might generated by different sets of rules. Both Markus and his supervisors had entered a steep learning curve, but it was fun and the report got a top grade and Markus later said it was this experience that later convinced him to apply to DTC and do a DPhil. Other students were extremely interested Markus' project: He was to give a 20 minutes presentation but it lasted 90 minutes since there was 70 minutes worth of questions!!


Markus Gerstel

In 2015 November I attended a Workshop at Lorentz Center in Leiden and it was obvious that the technique of statistical alignment were ideally suited for word alignment and should also be able to de cognate recognition [the analogue of sequence homology]. The problems and advantages are very similar: you are not forced to consider only one alignment and can do parameter testing for both sound change [subsitution] and insertions-deletions. There are also differences, where words are more complicated: i. there are correlations within a word and ii. the evolution of different words are correlated. One advantage is that words are much shorter than sequences so more advanced models becomes feasible. Over the last years the amount of data on words in different languages has increased significantly. This is not funded at the same level as bioinformatics, but the sitution is getting a lot better.


Søren Wichman


Gerhard Jaeger

My attendance of this workshop comes from my habit of always speaking to the people I sit next to on the plane: You always learn something, either about your destination or the person. On one flight I sat next to a Søren Wichman and it emerged that we had gone in the same Kindergarten and both were interested in Historical Linguistics and he worked at an Institute in Leipzig headed by Bernhard Comrie. At the conference I also discussed a lot with Gerhard Jaeger and Johannes Wahle and we hope to move statistical alignment forward in historical linguistics.


Johannes Wahle