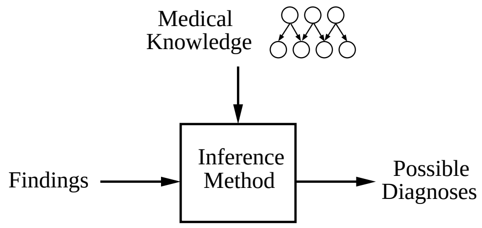
Lecture 24:

Clinical and Bioinformatics Applications

Sam Roweis

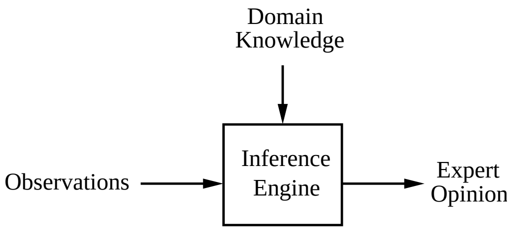
March 31, 2003

Medical Diagnosis



- In medical diagnosis the observations are clinical findings (what's wrong with this person), the domain knowledge represents which diseases or ailments have which symptoms, as well as which diseases are most likely for certain types of patients.
- The expert opinion takes the form of possible diagnosis (what ailment is most likely to be causing their problems).

General Motivation for Expert Systems

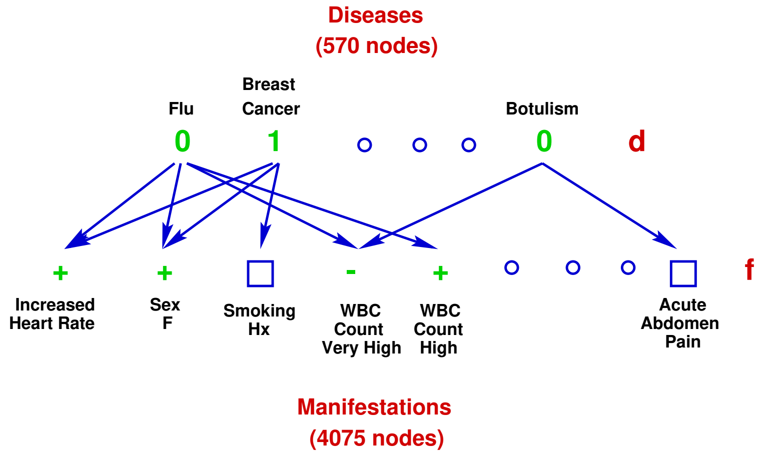


- Expert Systems attempt to combine domain knowledge with noisy observations and use a rational inference engine (often probabilistic) to come up with a conclusion or opinion.
- The two main problems in expert systems are how to encode the domain knowledge and how to perform inference efficiently.

Quick Medical Reference (QMR-DT)

- Quick Medical Reference, Decision Theoretic (QMR-DT)
  Is a very large graphical model based on expert knowledge acquired from medical doctors and clinical records in hospitals.
- There are 570 diseases and 4075 manifestations, which include symptoms, demographic data about the patient, medical history, and results of laboratory tests.
- We represent these using binary random variables d_k and f_i, encoding all non-binary manifestations (e.g. continuous values or categorical findings) with one-hot or range values.
- The domain knowledge was *not* learned from data directly using maximum likelihood, etc. Instead it was captured from the historical medical literature and from expert opinions and encoded into the graphical model by hand.

Quick Medical Reference (QMR-DT)



Inference is the Key

- The full posterior is huge (exponential in the number of diseases), so we can only ever hope to compute its marginals.
- Even just to compute the likelihood requires a large amount of work because we have to sum over all possible disease configurations.

d: disease configuration

f: findings

prior conditional joint

$$\underbrace{P(d\,|\,f)}_{\text{posterior}} = \frac{\overbrace{P(d)}^{\text{prior}}\ \overbrace{P(f\,|\,d)}^{\text{conditional}}}{\sum_{d'} P(d')\ P(f\,|\,d')}\qquad \frac{\overbrace{P(d,f)}^{\text{joint}}}{\underbrace{P(f)}_{\text{probability of evidence}}}$$

posterior

probability of evidence

Quick Medical Reference (QMR-DT)

- The graphical model asserts that manifestations are conditionally independent given the diseases:

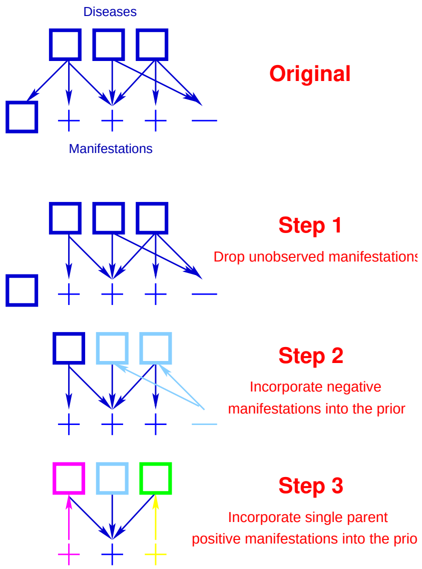
$$p(\mathbf{d},\mathbf{f}) = \left[\prod_k p(d_k)\right]\left[\prod_i p(f_i|d)\right]$$

- The conditional model for activation of the manifestations given the diseases is noisy-OR:

$$\log p(f_i = 0|\mathbf{d}) = w_{i0} + \sum_k w_{ik} d_k$$

- Most of the time very few diseases are active (less than 9), and zero or one diseases account for 72% of the mass under the disease prior.
- Also, usually between a few and a hundred manifestations are observed out of the 4075 possibilities.
- The noisy-OR weights are also very sparse: only 2% are nonzero.

Inference in 2-layer binary noisy-OR networks



But there is a trick...

- The Quickscore Algorithm (Heckerman 1989) computes $P(f)$ in time exponential in number of positive findings with multiple parents.
- The trick is that negative findings and positive findings with only one parent can be absorbed into the prior.
- Still, with 100 observations we would still have to sum over $2^{1}00$ configurations.

Approximate Inference

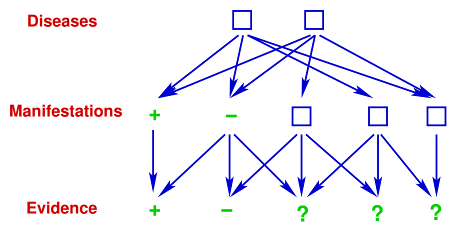
- Even with the quickscore trick, exact inference is often intractable in networks as large as QMR-DT.
- So practitioners resort to approximate inference methods which attempt to *estimate* the marginals $p(d_k|\mathbf{f})$ rather than computing them exactly.
- This is a large and complex area of research, but essential to making QMR a practical diagnosis system.

Discrete Sequences in Computational Biology

- There has recently been a great interest in applying probabilistic models to analyzing discrete sequence data in molecular and computational biology.
- There are two major sources of such data:
  - amino acid sequences for protein analysis
  - base-pair sequences for genetic analysis
- The sequences are sometimes annotated by other labels, e.g. species, mutation/disease type, gender, race, etc.
- Lots of interesting applications:
  - whole genome shotgun sequence fragment assembly
  - multiple alignment of conserved sequences
  - splice site detection
  - inferring phylogenetic trees

QMR-DT Observation Process

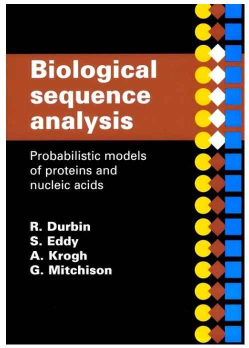
- Other issues: there should be a distinction between *unobserved* manifestations and *observed negative* manifestations.
- Observation is not independent of result: doctor's do the tests they expect will give them important info.
- Modeling this observation process is key to using QMR in practice (see work of Quaid Morris).



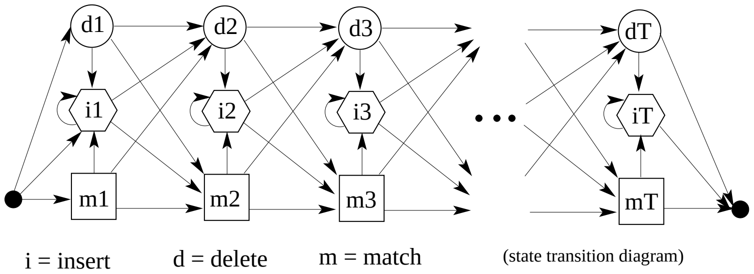
Main Tool: Hidden Markov Models

- HMMs and related models (e.g. profile HMMs) have been the major tool used in biological sequence analysis and alignment.
- The basic dynamic programming algorithms can be improved in special cases to make them more efficient in time or memory.

See the excellent book by Durbin, Eddy, Krogh, Mitchison for lots of practical details on applications and implementations.



Profile (String-Edit) HMMs

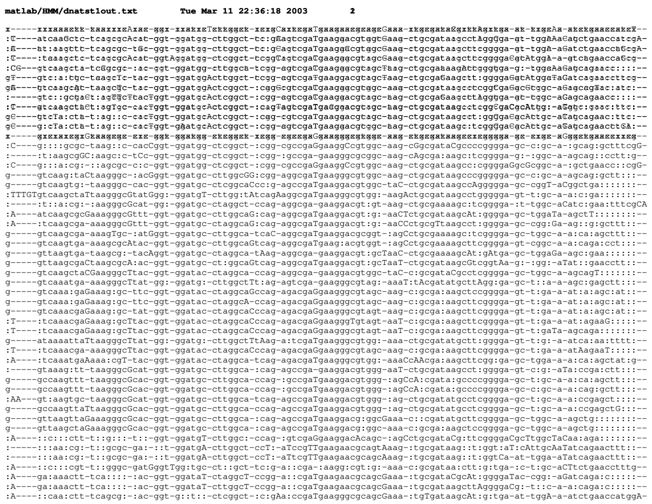


- A "profile HMM" or "string-edit" HMM is used for probabilistically matching an observed input string to a stored template pattern with possible insertions and deletions.
- Three kinds of states: match, insert, delete.
  m_j – use position j in the template to match an observed symbol
  i_j – insert extra symbol(s) observations after template position j
  d_j – delete (skip) template position j

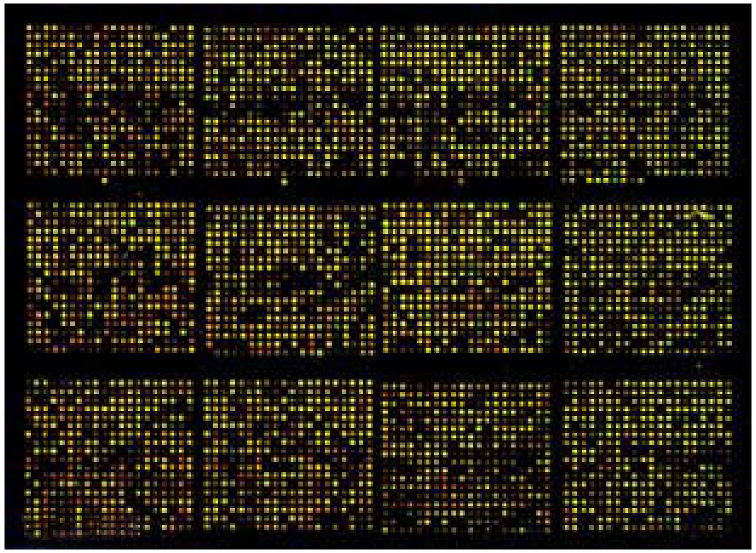
Microarray Data

- A final source of clinical data now gaining attention for analysis by probabilistic graphical models is data generated from DNA microarry experiments.
- A "DNA chip" is manufactured with several banks of "probe sequences" attached to the surface in known location.
- A test solution is washed over the chip, and species in the test solution having a high affinity with the probe sequences are more likely to bind to the probes, or stay bound for longer.
- This binding is measured using optical flourescence or electrical conductivity signals, giving a signal which tells us how "similar" the probe sequence was to some subsequences in the test solution.
- By repeating this experiment over multiple test solutions representing different mutants or individuals we can generate an enormous quantity of continuous measurement data.

Example


matlab/HMM/dnatstlout.txt Tue Mar 11 22:36:18 2003 2

Microarray Data



Analysis of Microarray Data

- The continuous measurements from the microarrays can be analyzed using many of the models we have studied, e.g.
  - factor analysis
  - mixture models for clustering
  - classification with naive bayes or logistic regression