Logistics

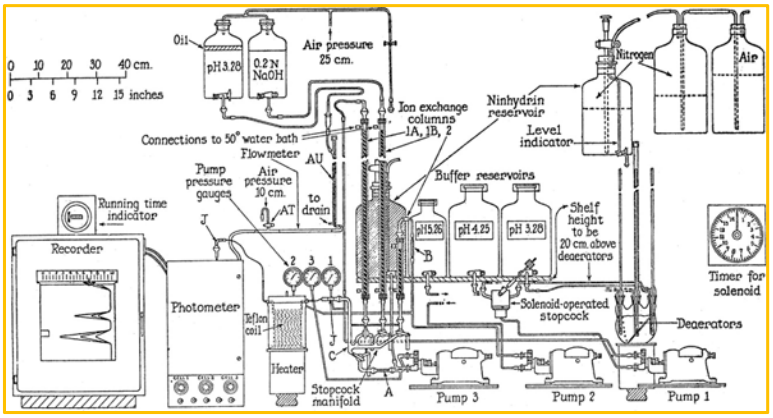
- Problem set 3: Due Friday, Sep 30
  - Nucleotide substitusion models
- Problem set 4: Due Friday, Oct 7
  - Log odds scoring
- IN CLASS EXAM Tuesday, Oct 11
  - Covers lectures 1-9, up to log-odds scoring
  - Closed book, 2 pages of notes
- 711-2 assignment Due Friday, October 14
- Mid term break Oct 17-21

Amino acid sequencing predates DNA sequencing

Sanger develops peptide sequencing based on electrophoresis and chromatography and wins 1958 Nobel prize for "for his work on the structure of proteins, especially that of insulin."

Trends in Biochem. Sci, 99

Protein Sequencing by Stepwise Degradation



Edman, P. (1950) Method for Determination of the Amino Acid Sequence in Peptides**.** Acta Chem. Scand., 4, 283–293

Spackman, D. H., Stein, W. H., and Moore, S. (1958) Automatic recording apparatus for use in the chromatography of amino acids. *Anal. Chem.* **30,** 1190–1206

A Beckman-Coulter Porton LF3000G protein sequencing machine

By Michael Pereckas from Milwaukee, WI, USA - Porton, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=4065582

Amino Acid Substitution Matrices

Overall strategy for both PAM and BLOSUM

1. Trusted amino acid alignments
2. Obtain amino acid pair counts $\left(A_{xy}^N\right)$with corrections for
   - Evolutionary divergence
   - Sample biases
3. Estimate substitution frequencies, q_{xy}^N, from pair counts, A_{xy}^N
4. Log odds substitution matrix:

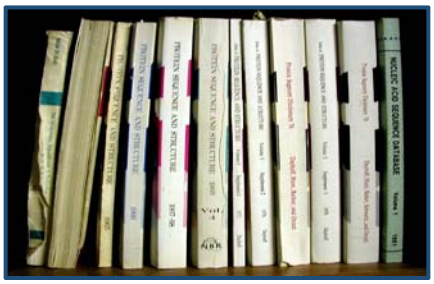
$$S[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$$

Frequency of *x* aligned with *y* in sequences with divergence *N*

Scaling constant

Frequency of *x* aligned with *y* in "random" sequences

Atlas of Protein Sequence & Structure 1965 - 1978



Examined 1572 changes in 71 groups of closely related proteins

Margaret Dayhoff

PhD in Chemistry, 47

Watson Computing Lab Fellow 47 - 48

PAM matrix training data

Examined 1572 changes in 71 groups of closely related proteins drawn from 34 protein superfamilies, e.g.:

```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

Glyceraldehyde 3-phosphate dehydrogenases