Secondary structure prediction

Andrew Torda, wintersemester 2008 / 2009, 00.904 Angewandte …

Is secondary structure prediction really important ?

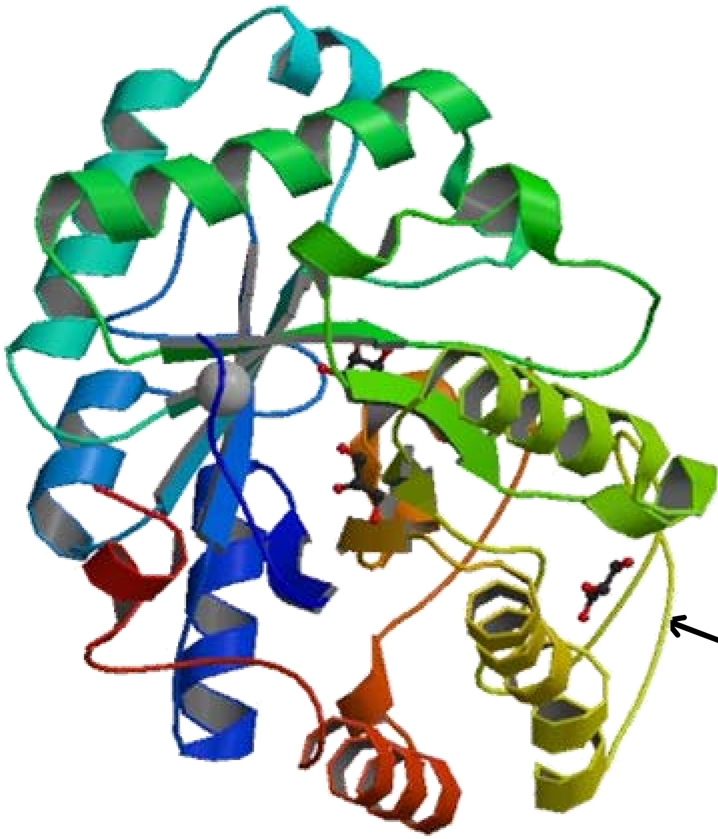
- not if we could do full structure prediction reliably

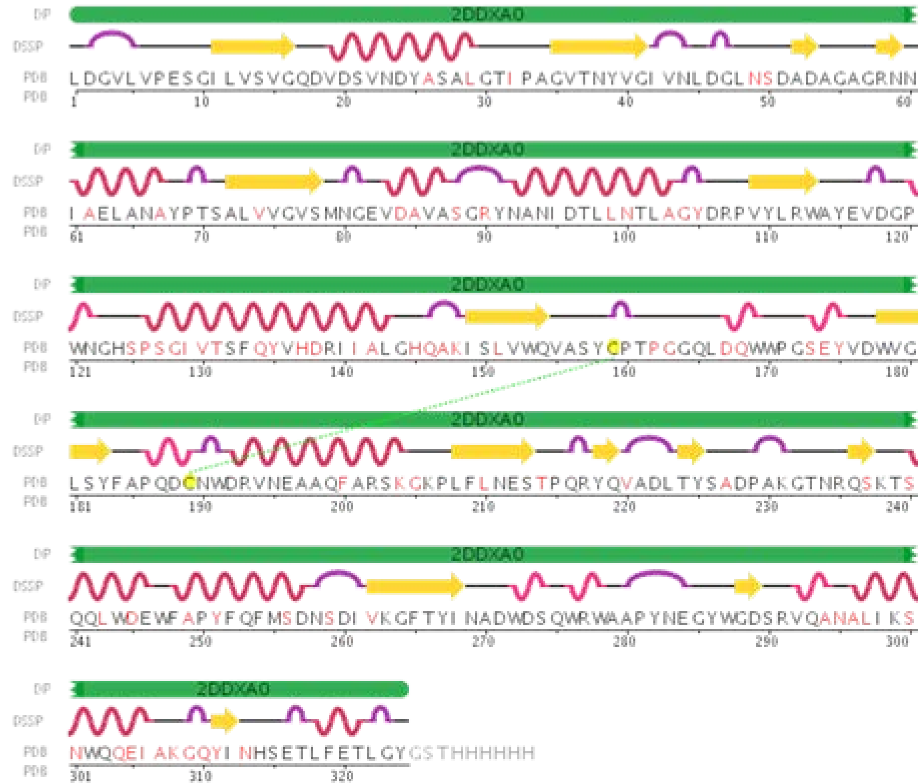
Why worry ?

Looks tempting...

The mission

- Go from ADADQRADSTR
- to HHH__EEEEHH





Looks easy

These lectures

- why do we care about secondary structure prediction ?
- history
- definitions
  - secondary structure
  - prediction accuracy
- neural nets
- neural nets for secondary structure
- other approaches

- Does Prof Torda like
  - secondary structure prediction ?
  - neural nets ?

Who cares about secondary structure prediction ?

- seems like an easier problem
- belief (1)
  - prediction of secondary structure
  - put these units together
  - easy protein structure prediction
- belief (2)
  - secondary structure forms first in protein folding
  - not proven - not necessarily true
- real evidence of statistical trends
- huge history
- very very popular in biological labs
- techniques might be applicable to other problems
  - predicting
    - solvent accessibility, coils, membrane bound

Why should secondary structure be predictable ?

There are statistical preferences

- obvious
  - alanine likes helices
  - proline does not like helices (no H-bond donor)
- less obvious
  - β-strands more likely to be buried
  - α-helices amphipathic
  - residues have preferences (hydrophobic, polar, charged..)
  - would expect predictable patterns

Hamburg Gesetze

Conventions – different names and types of secondary structure

detailed	condensed		
H	H	α-helix	most important
E	E	β-strand	
B		β-bridge	
G	H	3-10 helix	
I	other / L / coil/…	5 helix	
T		H bonded turn	
S		bend	

We will mostly stick to H, E, other (coil)

A Trottelvorhersage

- take set of representative proteins
- assign secondary structure
- count number of times residue occurs in each type

A better predictor

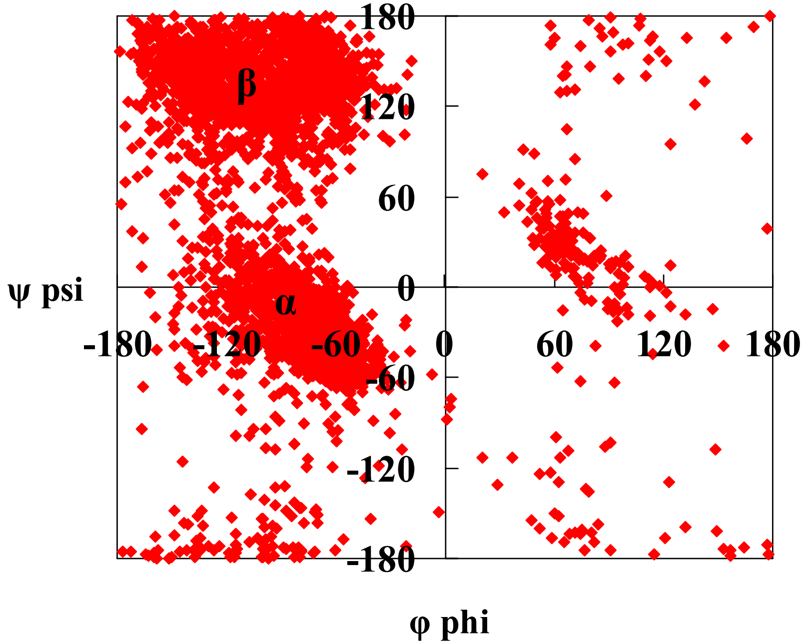
- You cannot have an α-helix of one residue
  - physically > 4 residues, usually more
  - EEE___HEE not possible
  - β-strands normally longer as well
- Chou and Fasman (1978)
  - look for stretches of 6 likely "H"
  - 5 likely "E" (β-strand)
- About 50-60 % correct

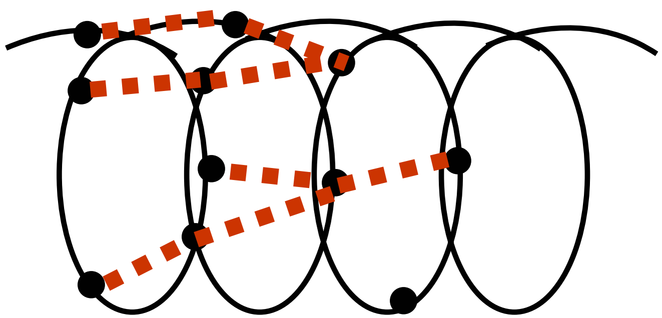
Defining secondary structure

Before going on, need some definitions

How rigorous is secondary structure ?

- defined by geometry or H-bonds ?

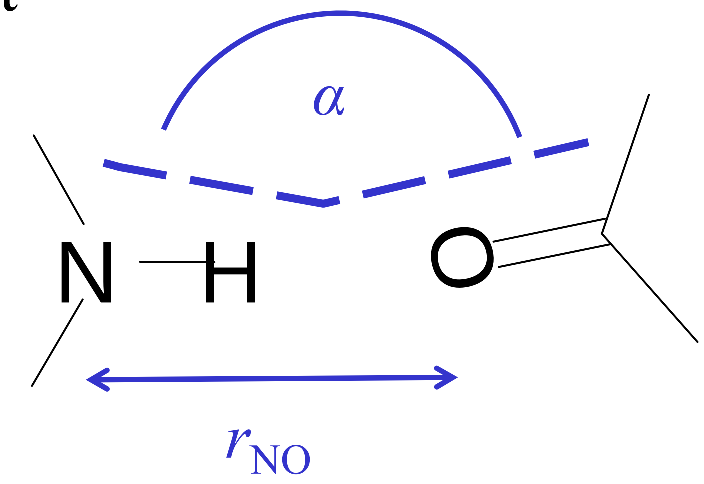




Maybe H-bonds are a bit better

How well is an H-bond defined ?

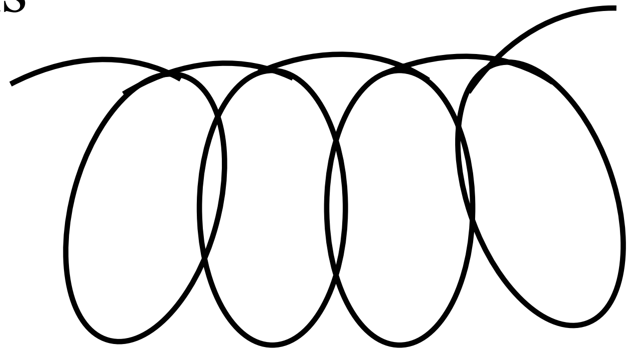
- H-bond is "in principle" well defined but
  - proteins have errors / are an average
  - not all geometry is ideal
  - not all H-bonds are the same
- Consequence
  - slight arbitrary element
  - how big is r_{NO} ?
  - how flat is α ?
- Different programs might differ
  - about H-bonds
  - about exact secondary structure



Different definitions of secondary structure

Assignments will differ between programs

- most differences at ends



Where will you meet this ?

- spdbv, rasmol, chimera…
- many programs for protein analysis

Most important ?

- DSSP (Kabsch and Sander)
  - pascal -> C (astonishingly ugly, grässlich, nicht robust)
  - free code, popular
- defines 8 types of secondary structure
- based on H-bond definition
- well described in paper

Kabsch, W & Sander, C, Biopolymers, 1983, 2577-2637, Dictionary of protein secondary structure

Measuring prediction accuracy Q3

- how many α-helical residues are correct ?
  - number of correct α-helix/number really α-helical

$$Q_{\alpha} = \frac{number\ residues\ correctly\ predicted\ as\ \alpha}{number\ residues\ observed\ as\ \alpha}$$

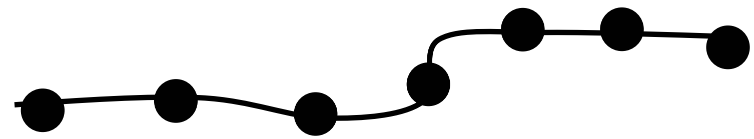
- more generally

$$Q_{3} = \frac{number\ residues\ correctly\ predicted}{number\ residues}$$

What is wrong with Q_3 ?

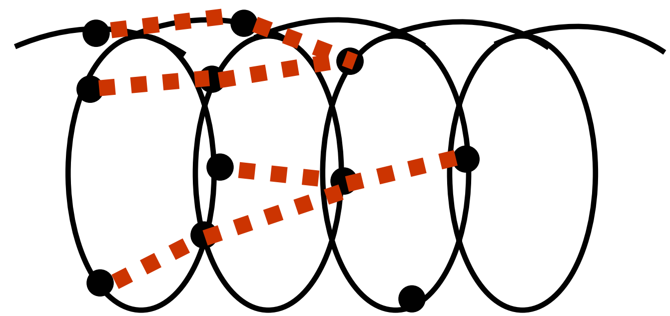
Not bad but

- EEEHEEEEE is a bit silly



Does not tell us about

- predicting
  - too much / too little
- different types of errors



Alternatives

- segment based (SOV)
- truth table
  - too hard

Generally use Q_3

		predicted		
		H	E	C
observed	H			
	E			
	C			

Baselines / Expectations

Proteins are

- 32 % α helix
- 21 % β strand/sheet
- 47 % others

Random guesses

- about Q_3 36 or 38 % correct

Approaches and history

Approaches / formulations

- statistics
  - most likely conformation of
    - an amino acid
    - a few amino acids
- information measures
  - how much does each position matter ?
  - how significant is an amino acid at some position ?
- rules
  - A followed by C three positions .. or a ...
  - automatic rule detection

Good review, Rost, B, 2002 http://cubic.bioc.columbia.edu/papers/2002_rev_dekker/paper.html

General philosophy

to predict this residue

A D S T S Q R A P P Q T A T Q R S E D R K K L W W

N_{res}

consider this window

Predict the conformation (H/E/?) of a residue based on his neighbours

- slide window along sequence
- N_{res} might be from 5 to 17

Garnier Osguthorpe Robier

Earliest somewhat successful approach

- Q_3 about 55 to 60 %
- N_{res} (window) = 17

Simplest approach

- look at residues in each conformation (α, β) in many proteins
  - make tables
  - not just which residues are present
  - which residues are most significant
- One side – information theory
- Others
  - log-odds probabilities

Why neural nets ?

There are statistical tendencies for amino acids to sec. struct

We expect some rules -examples

- residues near centre are important
- patterns ?
  - maybe if every fourth residue has some property = helix
  - alternating residues = β ?

- Simple neural nets are one way to pick up rules

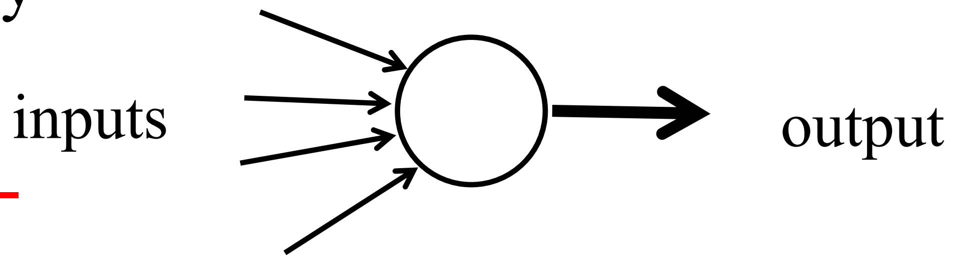
Neural nets...

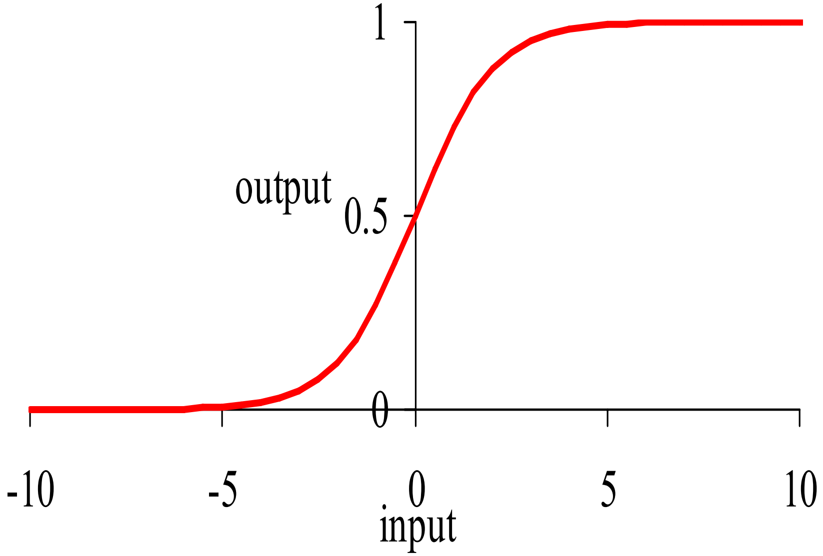
Many kinds

- soft computing lectures (Prof Stiehl)

Ours

- "feed forward / backwards propagation"
- one unit
  - switches off and on quickly



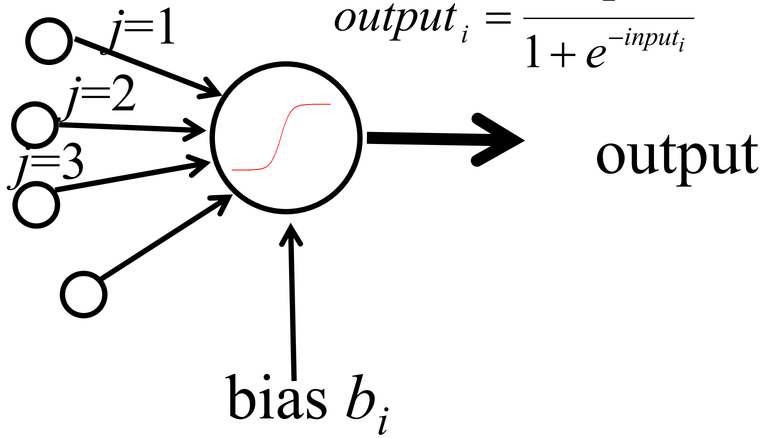


One unit of a net

- one unit sums up inputs and makes a decision (on / off)
- summing

$$output_i = \frac{1}{1+e^{-input_i}}$$

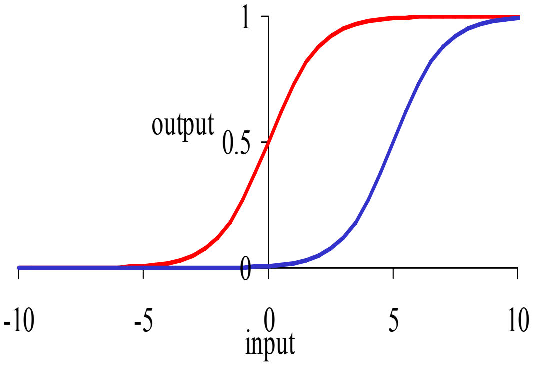
$$input = \sum_j W_{ij} + b_i$$



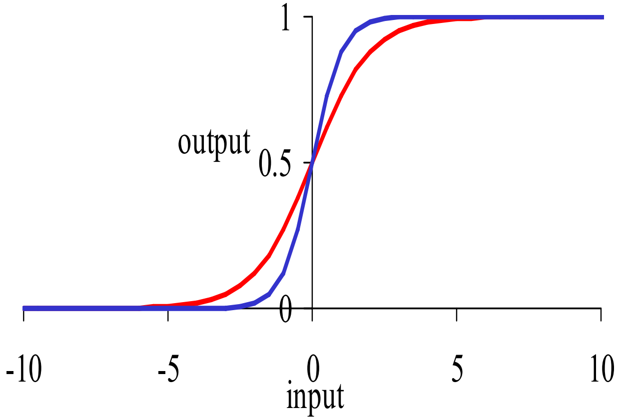
- what can we do to make it more interesting ?

Weights and biases

- bias moves left and right

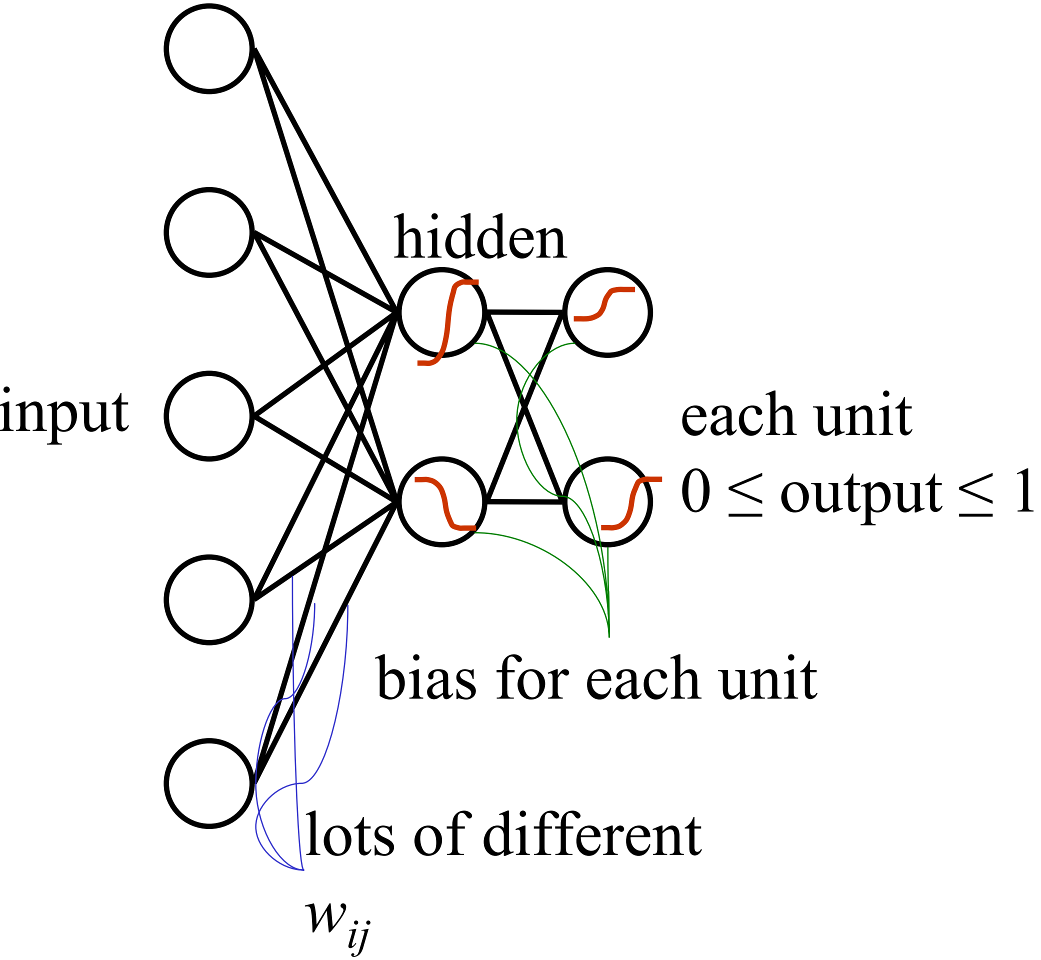


- our *w*'s make the curve sharp or flat
- a single unit may
  - respond quickly, slowly
  - be sensitive to some inputs
  - not care about others



bias übersetzung drama ! Abneigung ? not here..

A full neural net



- lots of weights
- lots of biases
- some "excitors", "inhibitors"

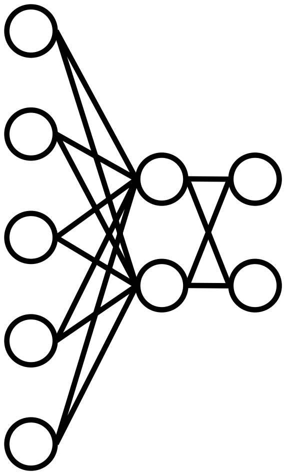
- should be possible to get some quite arbitrary output
- like coding up rules

What can one do ?

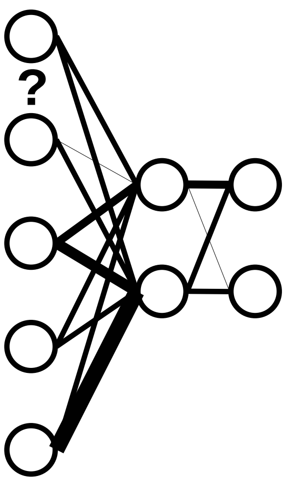
- get input into some reasonable form
  - set of 0's and 1's (good)
  - set of numbers in some controlled range
- very general mapping of input to some output
- how to get weights and biases ?
  - training

Training a net

- collect data
  - input data + matching output
- random weights and biases



```
while (not happy)
  show next pattern
  calculate output
  for each output node
     calculate (desired – observed)
     should we make a weight bigger or smaller ?
     small adjustment of weights
```

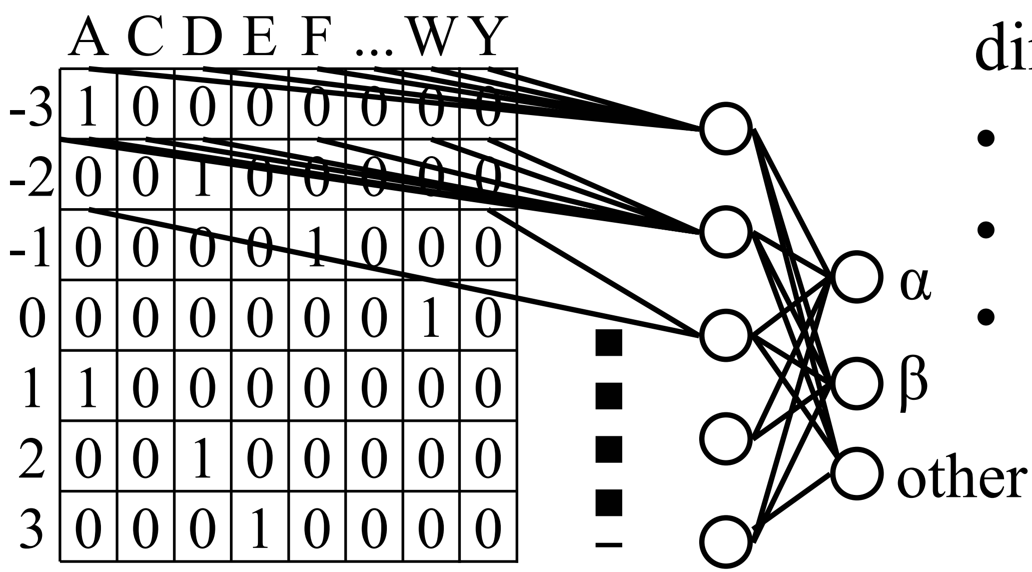


Over time

- weights and biases move up, down...
- hopefully becoming better

Neural Nets for secondary structure prediction

- input pattern
  - our central residue + neighbours ADADFWADER
- output
  - measured secondary structure HHH__EEEEH



dimensions

- at least N_{res} x 20 input nodes
- handful of hidden units
- about 3 output units

Earliest neural nets for secondary structure

- windows typically $13 \leq N_{res} \leq 19$
- hidden layer 5 < N nodes < 100
- output about 3 nodes

Success

- about Q_3 50 to 60 %
- Is this OK ?
  - not enough to build structures
- Q_β usually worse
- not much use

Where next ? Big change

Use of alignments

- consider one sequence and related neighbours
- and align
- get out average residue at each position
- Instead of binary (0 / 1) inputs, use the average at each position
  - 4/7 Leu, 1/7 Val, 1/7 Ile, 1/7 Ala
- why is this good ?
  - look at unusual "A" in row 2
    - is it significant ?
    - profiles average over weirdness
- averaging obvious, but there is more information

```
L D D Q R A   S T R
L D A Q R A D S T R
V D D Q R A W S T R
A D D Q R A A S S K
I D D Q R A D S T R
L D D Q R A G S T K
L D D Q R A C S T R
```

More information from alignments

- Alignment tells us
  - what is average residue type
  - how much does the residue vary
    - degree conservation
- Why should it matter ?
- Dogma
  - most mutations are bad, some very bad
  - buried regions are conserved
  - secondary structure is conserved
  - simple conservation is important
- Noise argument
  - predictions have random errors
  - think random errors, drunk walks

```
L D D Q R A   S T R
L D A Q R A D S T R
V D D Q R A W S T R
A D D Q R A A S S K
I D D Q R A D S T R
L D D Q R A G S T K
L D D Q R A C S T R
```

More information for each site

- 20 residues (0.0 to 1.0) x N_{res}
- deletion could be like a 21st residue
- how conserved is the central site ? (turn into a value 0 to 1)
- the other sites ? (turn into a value 0 to 1)
- now 22 inputs per site in window
- how to handle ends ?
  - add another kind of residue

Information for whole window

- overall composition (20 nodes ?)
- length of chain (small proteins are weird)

State of the art predictors

- Success ?
  - 72 to 77 %
  - β-strand no worse than α-helix (earlier a problem)
- all use sequence profiles
- somehow include preference for intact segments (H is more likely next to H)
  - extra layers / nets
- measures of reliability

Why this success ?

- neural nets have NOT improved
  - experience with training and details
- profiles, multiple sequences
- database growth

Warum sind neural nets hässlich ?

Can I see what has happened ?

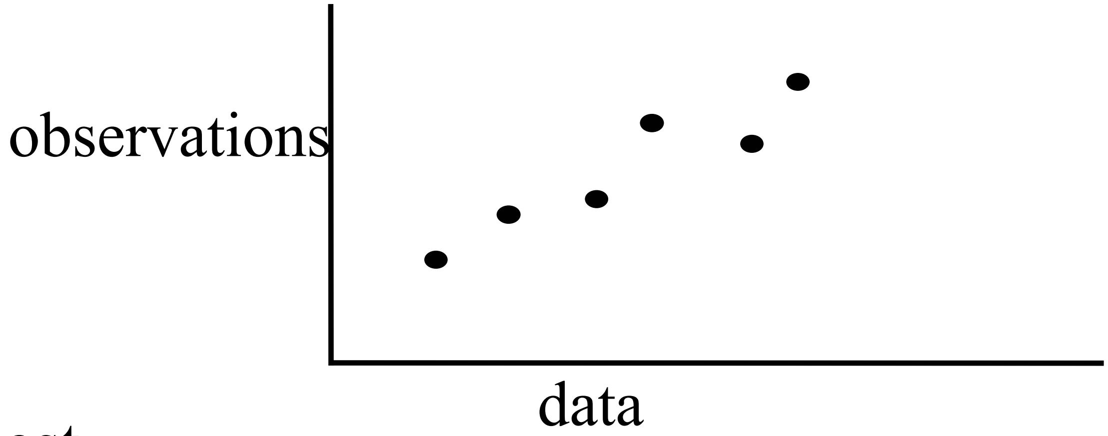
- can I work out the rules that turn on the α-helix unit ?

Number of variables

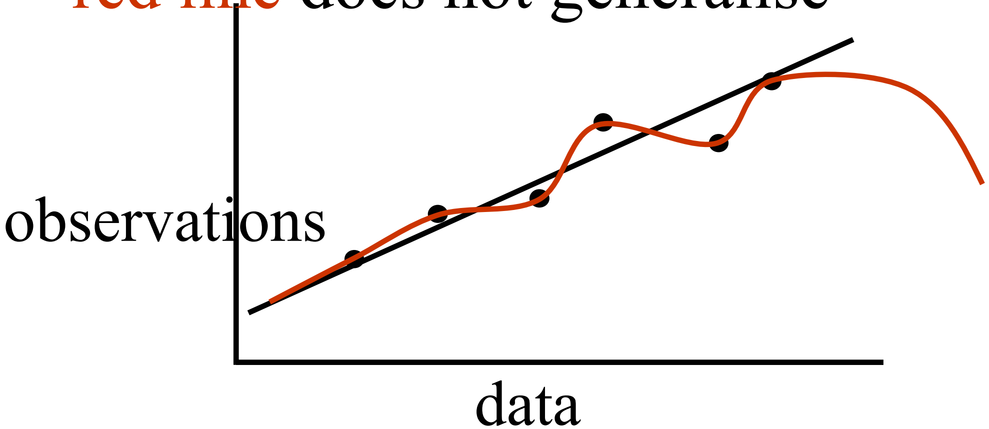
- weights + biases
  - typical 1000 to 50 000
- how many do I need ?
- are the extras harmless ?
  - recall vs. generalisation
  - too many connections
    - "fitting to noise"

Fitting to noise

- what is the best explanation of data ?



- red line fits data best
- black line is underlying model
- details are noise
- red line does not generalise



- best model
  - represents underlying behaviour
  - fewest parameters

Other learning / classifying procedures

- Belief and aim
  - secondary structure is a property of a residue and its neighbours
  - any procedure which maps

```
ADADQRADSTR
    ↓
HHH__EEEEHH
```

- any idea from
  - statistics
  - pattern discovery
  - classification
  - decision tree construction
  - hidden Markov models
  - support vector machines..

Limits

Regardless of method

- If we have coordinates, no consensus as to secondary structure !
  - limit could be 88 %

All current methods limited to common proteins

- best on soluble, globular proteins

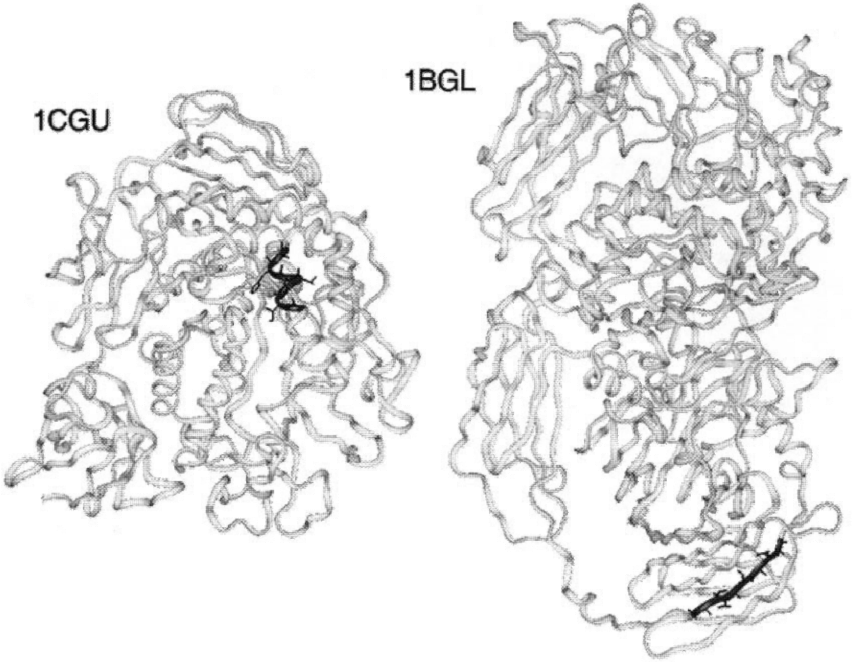
Real limit lower

- trying to predict conformation from local properties
- is is really a local property ?
- would you expect a pentamer defines local structure ?

(these are kind of things I like for exam questions)

Pentamers in different conformations

- can one really hope to predict secondary structure based on sequence ?
- first examples
  - search PDB and look at 5-mers (pentamers)
    - often same sequence in different conformation
- later 7-mers

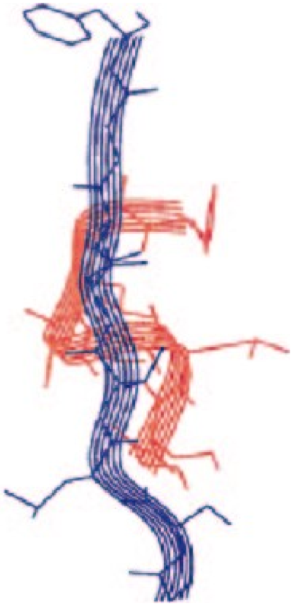


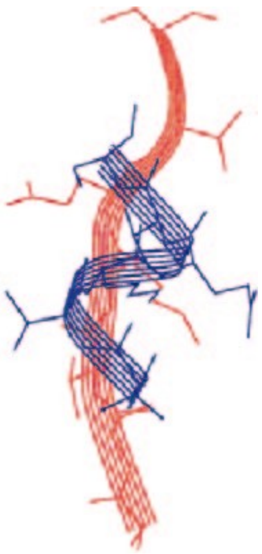
Mezei, M, Prot. Eng, 11, 411-414 (1998) Chameleon sequences in the PDB

even worse

- 8-mer pair, 1pht and 1wbc

7-mer pair, 1amp and 1gky

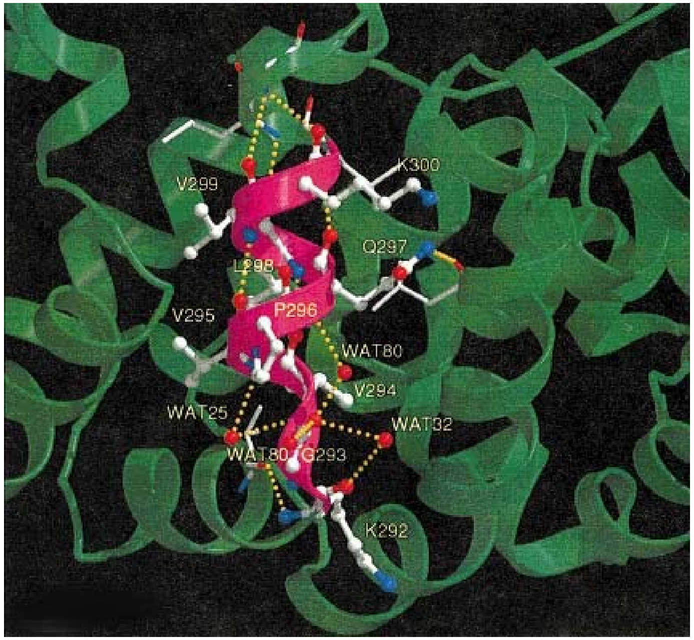




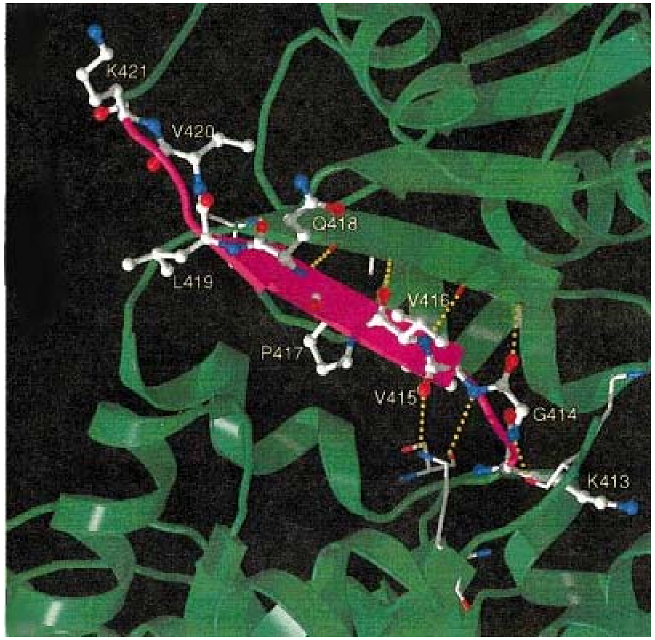
from Sudarsanam, S, Proteins 30, 228-231 (1998), "… identical octapeptides can have different conformations"

even worse

- 9-mer



1ial



1pky

- sequence `KGVVPQLVK` from two proteins

Zhou, X, Alber, F, Folkers, G, Gonnet,G.H., Chelvanayagam, G. Proteins, 41, 248-256 (2000) An Analysis of the Helix-to-Strand Transition Between Peptides With Identical Sequence

Minor and Kim (much worse)

- Take IgG-binding domain of protein G
  - write down an 11-mer
  - insert in one place
    - forms α-helix
  - insert in another
    - forms β-strand

A conclusion

- Secondary structure is largely determined by local effects
- secondary structure is very influenced by context / environment

Minor Jr, DL & Kim PS Nature, 380, 730-732 (1996) Context-dependent secondary structure formation of a designed protein sequence

Why spend all this time on neural nets ?

- Neural nets are most popular approach
  - secondary structure can be used towards full structure
- Underlying physics not well known
  - number of parameters totally empirical
- Lots of literature on neural nets
- Methods more generally applicable
  - rules might exist
  - not well understood / not well known
    - can we recognise a membrane bound piece of sequence ?
      - maybe it is a hydrophobic core
    - can we recognise sites for chemical modification
      - phosphorylation, acetylation, glycosylation... ?
- Neural nets could be useful for these