Cranfield UNIVERSITY

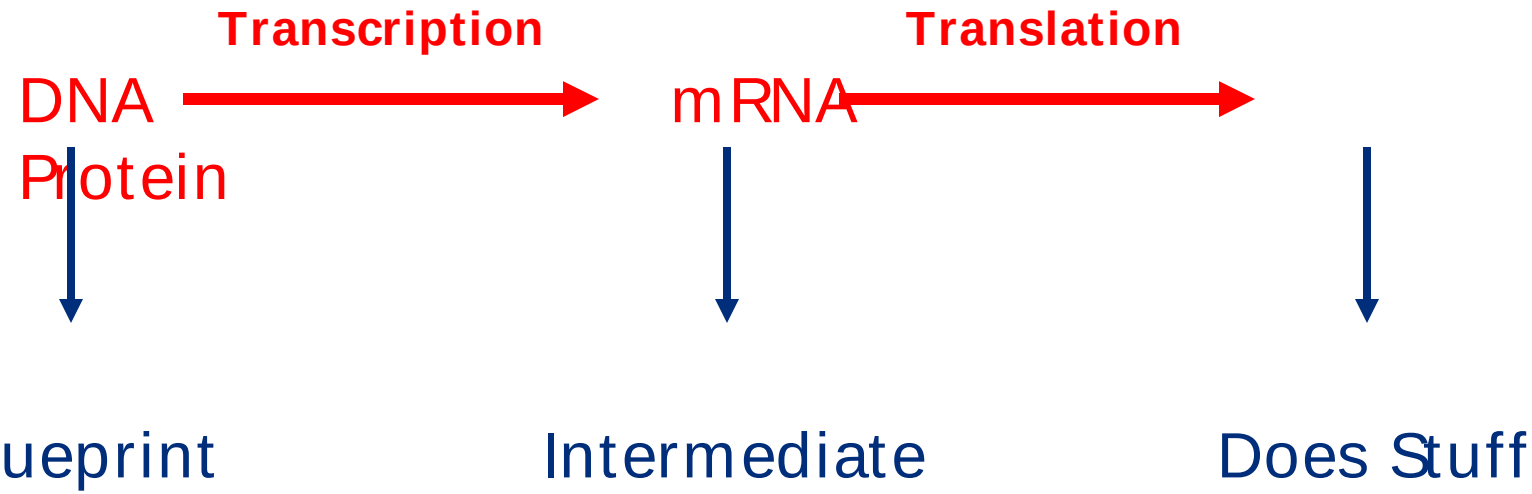
Perl: Doing something useful with the secrets of life

Structural Bioinformatics

Dr Lee Larcombe

MK Perl Mongers Technical Meeting 10/10/06

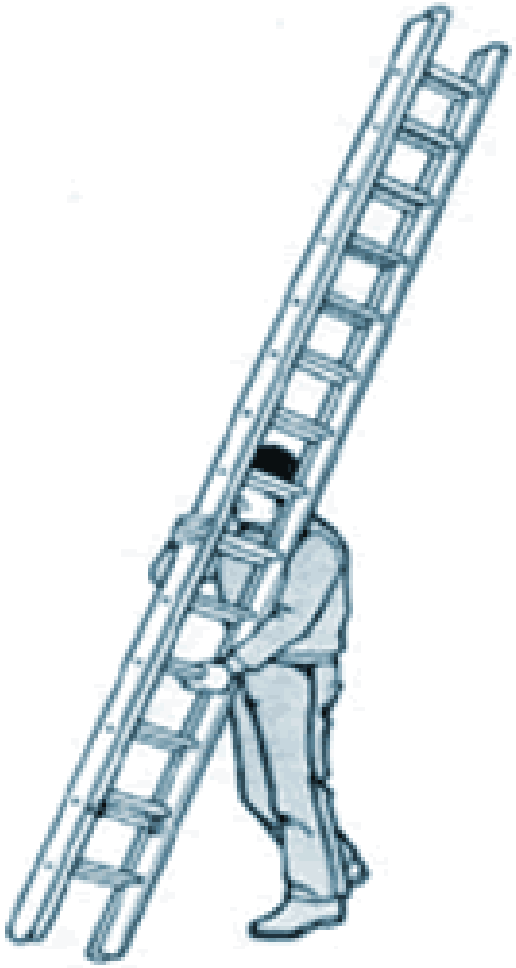
The Central Dogma of Biology




Gingerbread Baby
Preheat oven to 375 degrees.
In a large bowl, sift and mix together
3 cups flour
1/4 teaspoon salt
1 teaspoon baking soda
1 tablespoon ginger
1 teaspoon cinnamon
1/4 teaspoon ground cloves
1/4 teaspoon ground nutmeg
In another bowl or mixer cream:
12 tablespoon (1 1/2 sticks) unsalted butter
3/4 cup brown sugar
1 egg
Stir the dry ingredients into the creamed mixture
Then add:
1/2 cup molasses
1 tablespoon vanilla
Let dough rest at least 2 hours, roll dough 1/4 inch thick and cut with a cutter Bake 7 to 10 minutes on a greased cookie sheet. Do not peek! Makes 2 dozen cookies.

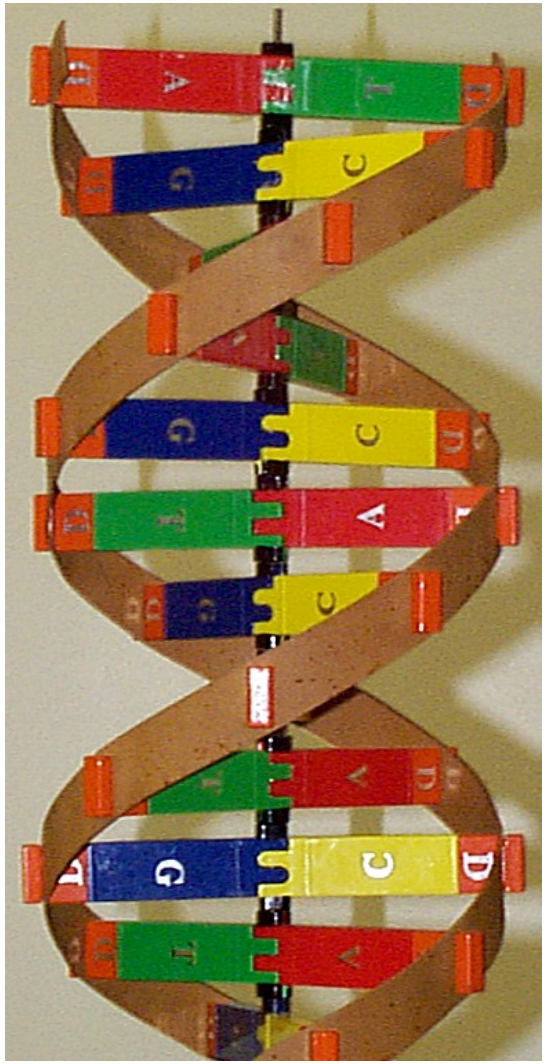
This is DNA

This is DNA

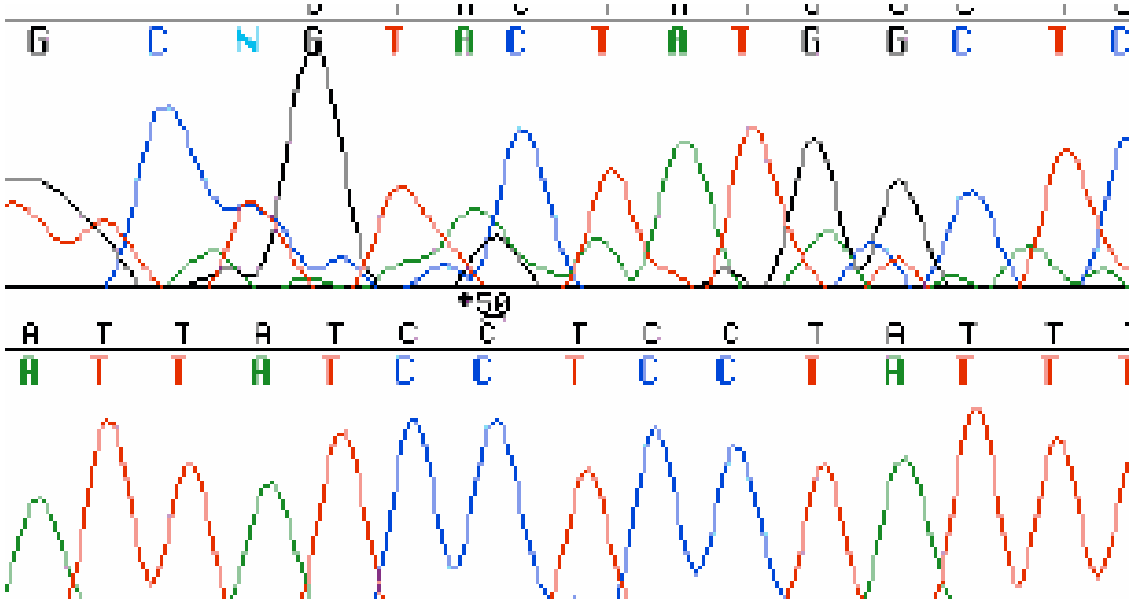


Cranfield UNIVERSITY

This is DNA



This is DNA



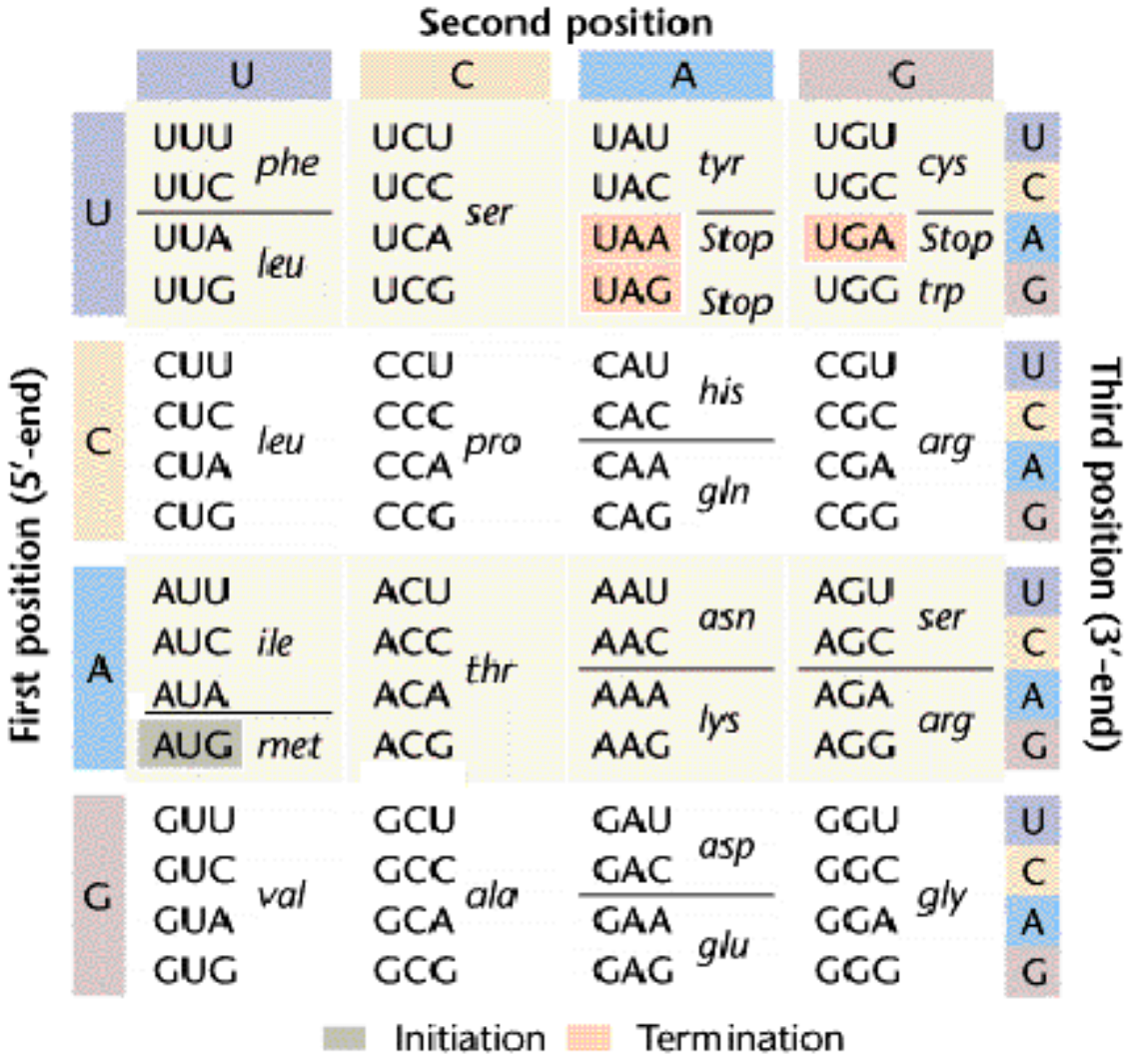

```
1   tcctggcatc agttactgtg ttgactcact cagtgttggg atcactcact ttccccctac
61  aggactcaga tctgggaggc aattaccttc ggagaaaaac gaataggaaa aactgaagtg
121 ttacttttt taaagctgct gaagtttgtt ggtttctcat tgtttttaag cctactggag
181 caataaagtt tgaagaactt ttaccaggtt ttttttatcg ctgccttgat atacactttt
241 caaaatgctt tggtgggaag aagtagagga ctgttatgaa agagaagatg ttcaaaagaa
301 aacattcaca aaatgggtaa atgcacaatt ttctaagttt gggaagcagc atattgagaa
361 cctcttcagt gacctacagg atgggaggcg cctcctagac ctcctcgaag gcctgacagg
421 gcaaaaactg ccaaaagaaa aaggatccac aagagttcat gccctgaaca atgtcaacaa
481 ggcactgcgg gttttgcaga acaataatgt tgatttagtg aatattggaa gtactgacat
541 cgtagatgga aatcataaac tgactcttgg tttgatttgg aatataatcc tccactggca
```

The Central Dogma as Perl !


```
$sequence =~tr/ACGT/ACGU/;
```

This is RNA

mRNA → Protein



Translation uses the famous...

'genetic code'

Cranfield UNIVERSITY

mRNA → Translation → Protein

```
%codons = (GCU => A, GCC => A, GCA => A, GCG => A,
           UGU => C, UGC => C,
           GAU => D, GAC => D,
           GAA => E, GAG => E,
           UUU => F, UUC => F,
           GGU => G, GGC => G, GGA => G, GGG => G,
           CAU => H, CAC => H,
           AUU => I, AUC => I, AUA => I,
           AAA => K, AAG => K,
           UUA => L, UUG => L, CUU => L, CUC => L, CUA => L, CUG => L,
           AUG => M,
           AAU => N, AAC => N,
           CCU => P, CCC => P, CCA => P, CCG => P,
           CAA => Q, CAG => Q,
           CGU => R, CGC => R, CGA => R, CGG => R, AGA => R, AGG => R,
           UCU => S, UCC => S, UCA => S, UCG => S, AGU => S, AGC => S,
           ACU => T, ACC => T, ACA => T, ACG => T,
           GUU => V, GUC => V, GUA => V, GUG => V,
           UGG => W,
           UAU => Y, UAC => Y,
           UAA => x, UAG => x, UGA => x,
          );
```

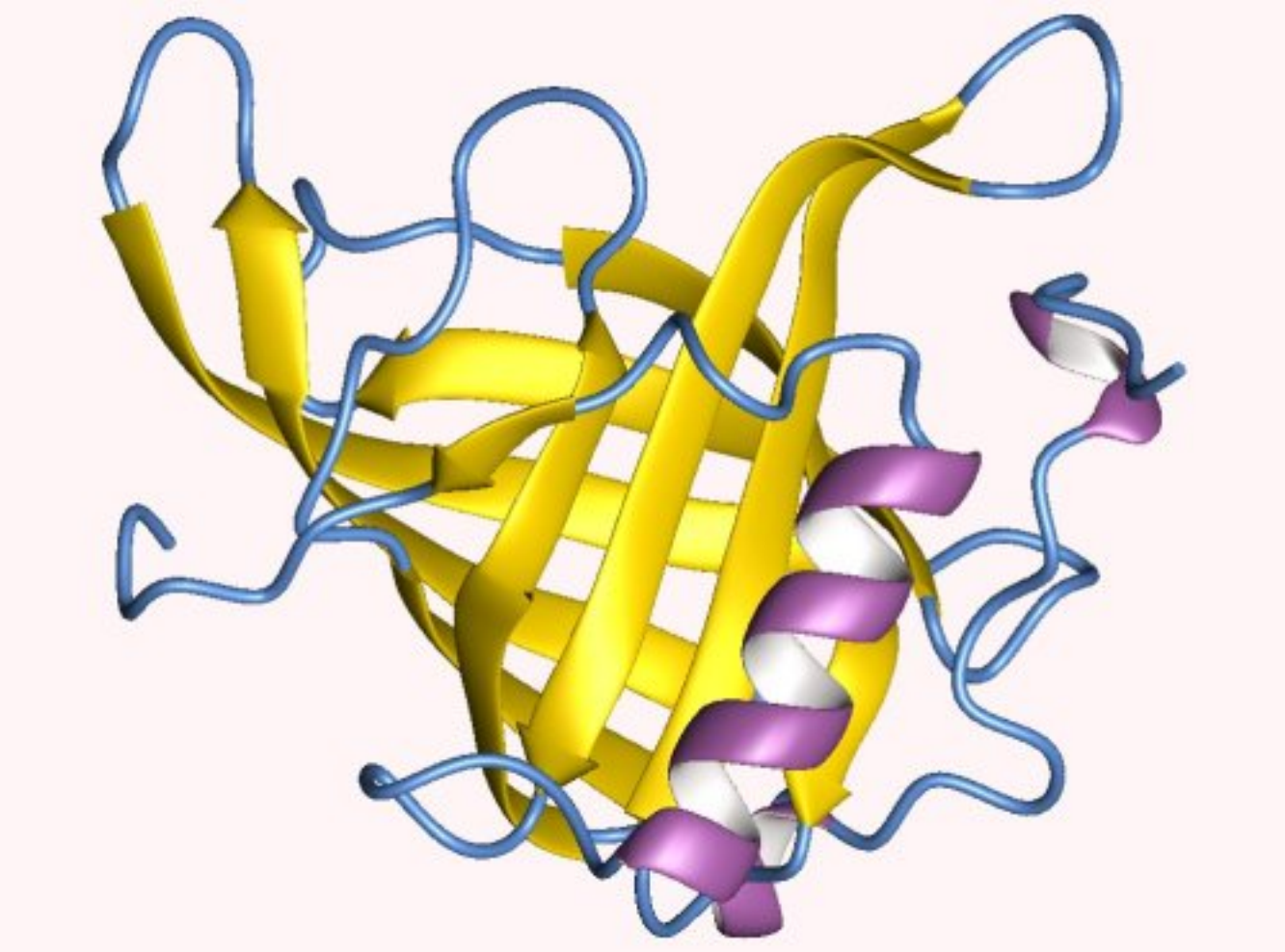
```
while ($rna=~s/(...)//){
            $protein = $protein.$codons{$1};
      }
```

This is Protein

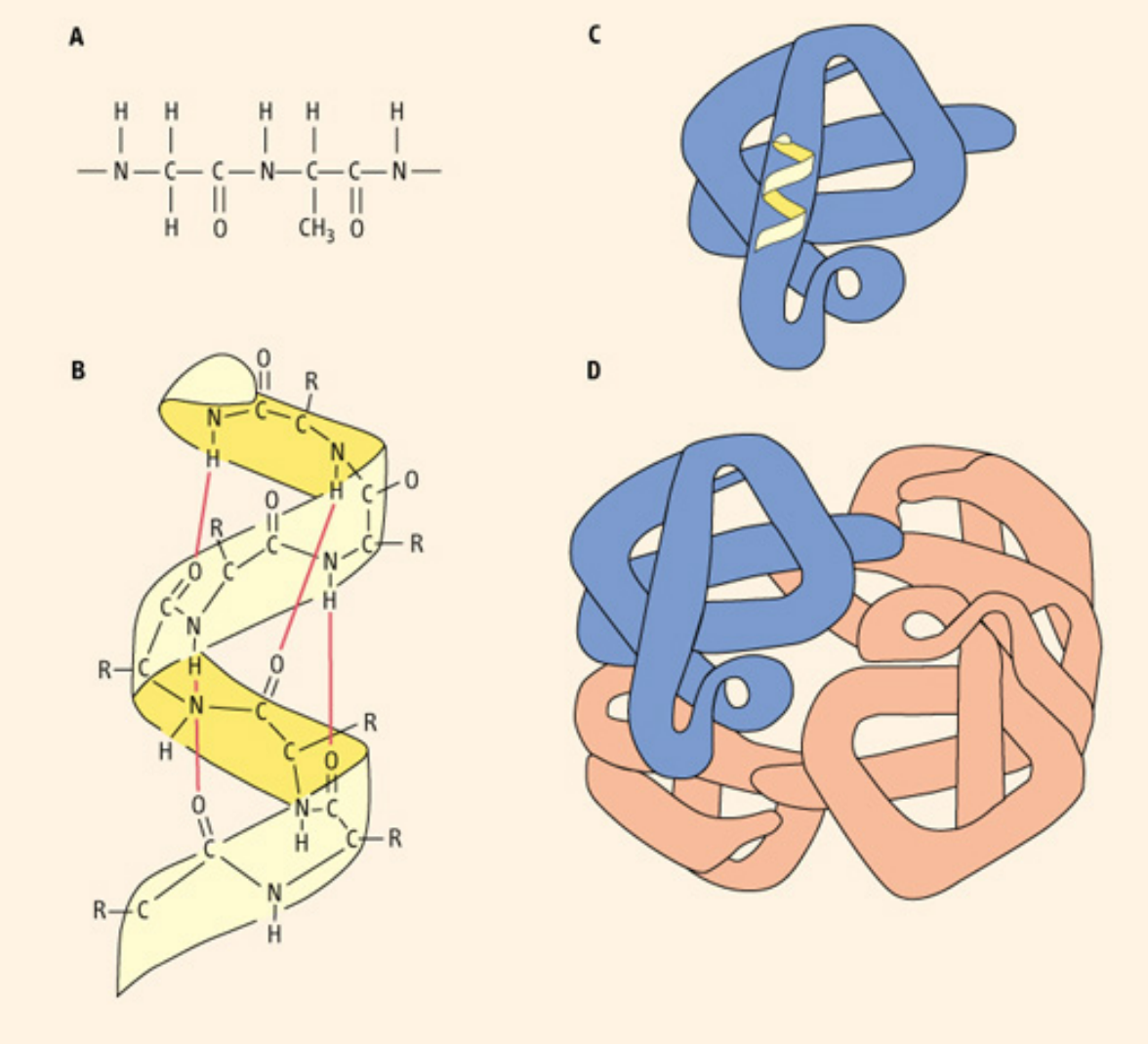
MTTPTL IVTPPSPPAP SYSANRVPQP SLMDKIKKIA
AIASLILIGT IGFLALLGHL VGFLIAPQIT IVLLALFIIS
LAGNALYLQK TANLHLYQDL QREVGSLKEI NFMLSVLQKE
FLHLSKEFAT TSKDLSAVSQ DFYSCLQGFR DNYKGFESLL
DEYKNSTEEM RKLFSQEIIA LKGSVASLRE EIRFLTPLAE
EVRRLAHNQQ SLTVVIEELK TIRDSLRDEI GQLSQLSKTL
TSQIALQRKE SSDLCSQIRE TLSSPRKSAS PSTKSS

This is Protein

This is Protein



Protein has 'levels' of structure



Knowing the structure is useful

Knowledge of structure can help us …

- **Design diagnostic tests**
- **Design drugs**
- **Understand diseases**
- **Produce Vaccines**

Very difficult to predict structure

The chemical bond between each subunit can take either of two orientations ...

```
    MTTPTL IVTPPSPPAP SYSANRVPQP SLMDKIKKIA
AIASLILIGT IGFLALLGHL VGFLIAPQIT IVLLALFIIS
LAGNALYLQK TANLHLYQDL QREVGSLKEI NFMLSVLQKE
FLHLSKEFAT TSKDLSAVSQ DFYSCLQGFR DNYKGFESLL
DEYKNSTEEM RKLFSQEIIA LKGSVASLRE EIRFLTPLAE
EVRRLAHNQQ SLTVVIEELK TIRDSLRDEI GQLSQLSKTL
TSQIALQRKE SSDLCSQIRE TLSSPRKSAS PSTKSS
```

There are 2^{272} possible structures for this protein - only 1 is is the real one.

We don't understand all the rules!

Computers are fast - but its still hard

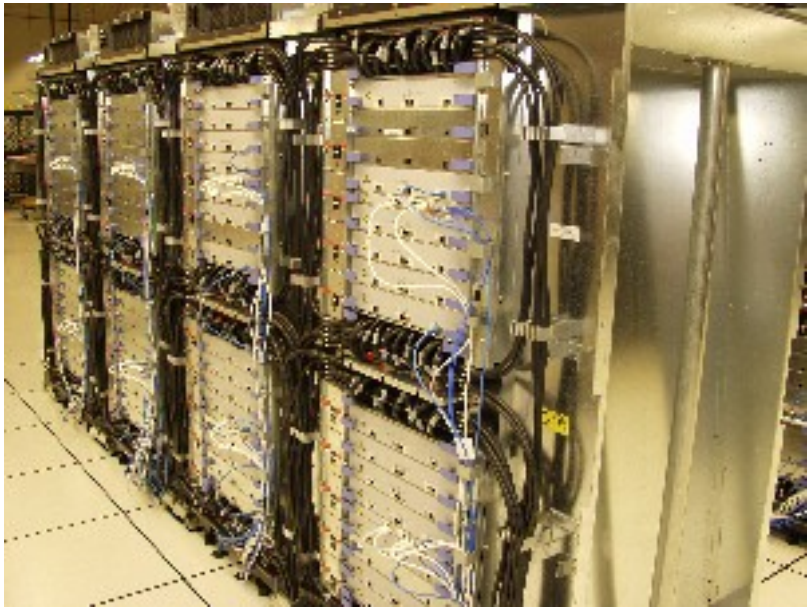
It's amazing that not only do proteins self-assemble -- fold -- but they do so amazingly quickly: some as fast as a millionth of a second. While this time is very fast on a person's timescale, it's remarkably long for computers to simulate.

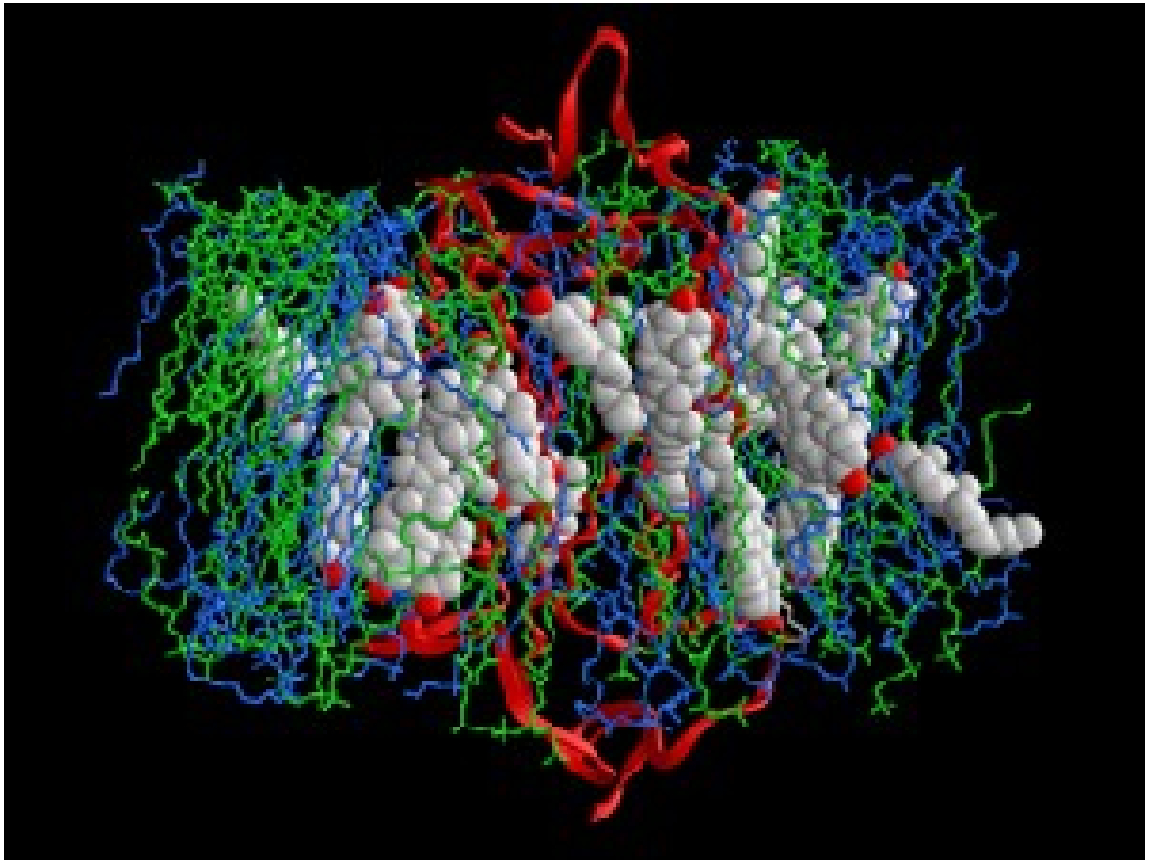
In fact, it takes about a day to simulate a nanosecond (1/1,000,000,000 of a second). Unfortunately, proteins fold on the tens of microsecond timescale (10,000 nanoseconds). Thus, it would take 10,000 CPU days to simulate folding -- i.e. it would take 30 CPU years! That's a long time to wait for one result!

Specifically designed by IBM to tackle problems in protein folding & structure optimisation

Reported cost > $100 million

Currently the world's fastest computer (360 teraflops)





G- protein coupled receptors in a membrane

(GPCRs)

represent more than half the current drug targets and a market of tens of billions of dollars annually

congestive heart failure, hypertension, stroke, cancer, ulcers, allergies, asthma, anxiety, psychosis, migraines, Parkinson's disease

Can prediction be simpler?

Can try to predict the 3D final structure - normally necessary to understand function

Or

Can just try and predict partial structure or surface from sequence to suggest targets for interaction

Antibodies

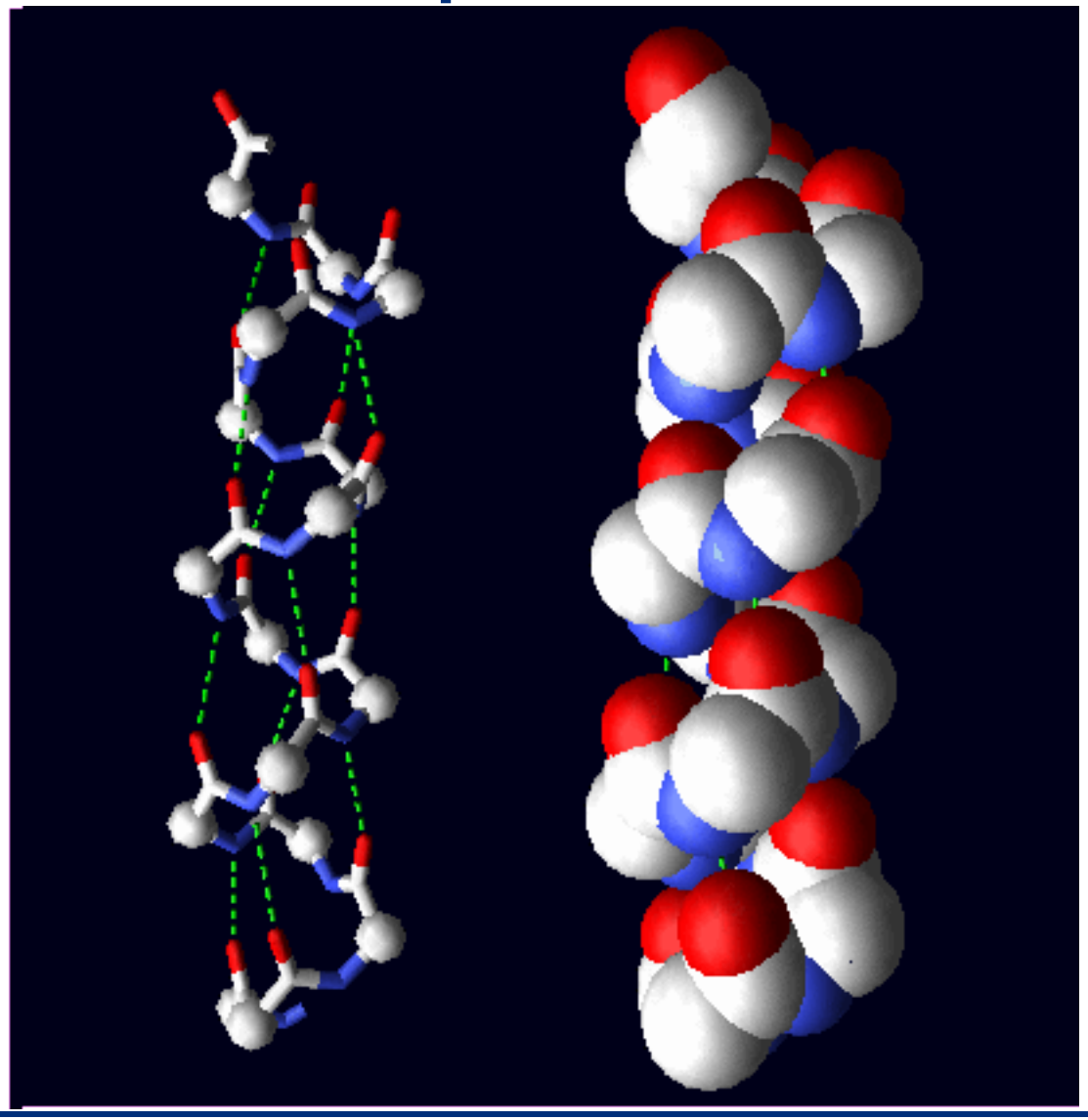
A simpler prediction

Basic methods of secondary structure prediction rely on statistical applications of 'propensity'

The propensity/ inclination/ tendency of protein subunit to be in a particular structure based on observation of known datasets

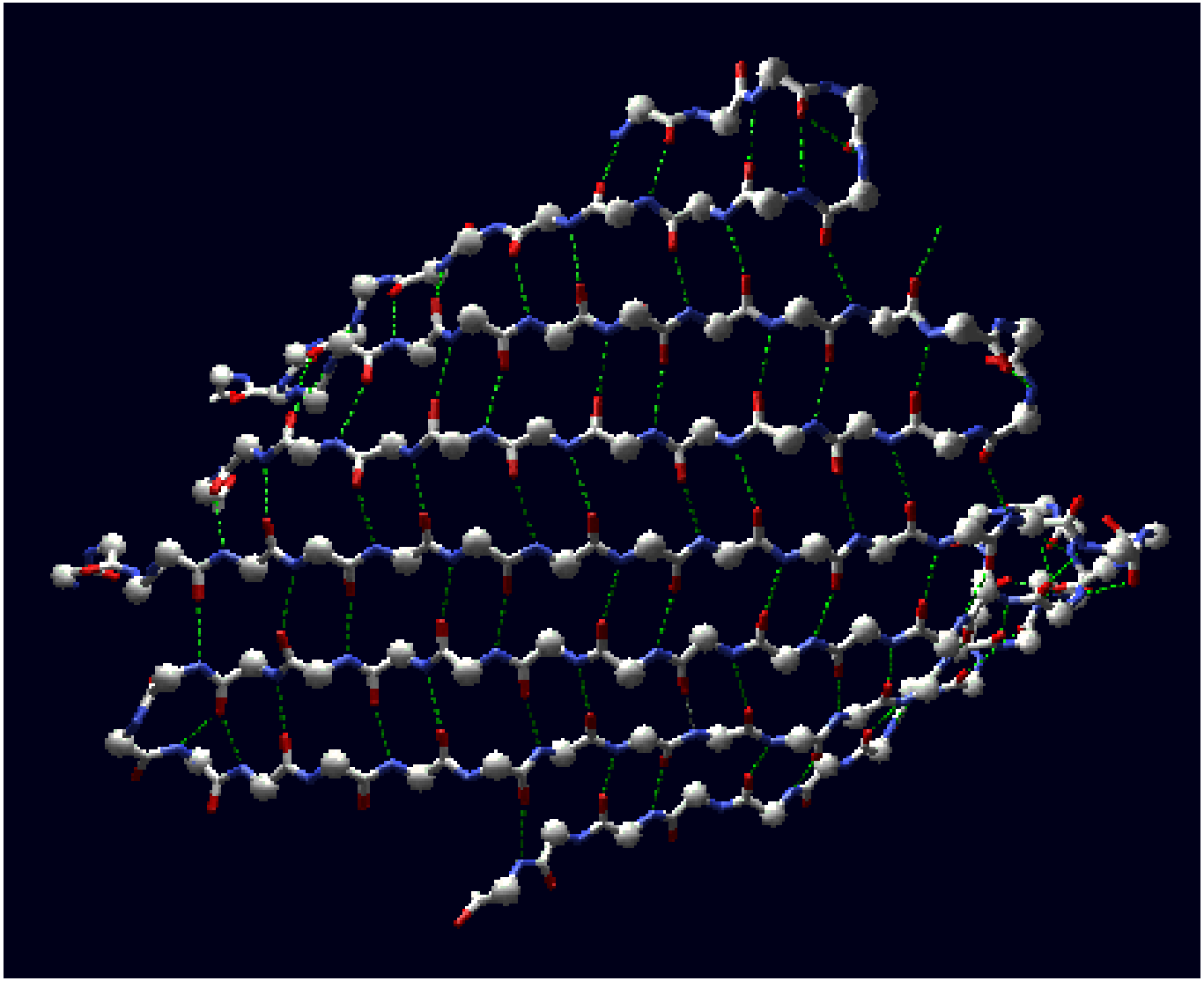
Cranfield UNIVERSITY

The Alpha Helix



Cranfield UNIVERSITY

The Beta Sheet



Propensity

$$P = \frac{n[I]^{[s]} / n[I]}{n^{[s]} / n}$$

P = propensity

I = subunit of interest

n[I] = number of subunits [I] in the database

n = total number of subunits in the database

$n[I]^{[s]}$ = number of subunits [I] in state of interest i.e. helices

$n^{[s]}$ = number of all subunits in the database in the state of interest.

Example

$$P^{[A]} = \frac{124 \ / \ 1640}{1246 \ / \ 10136} = 0.61$$

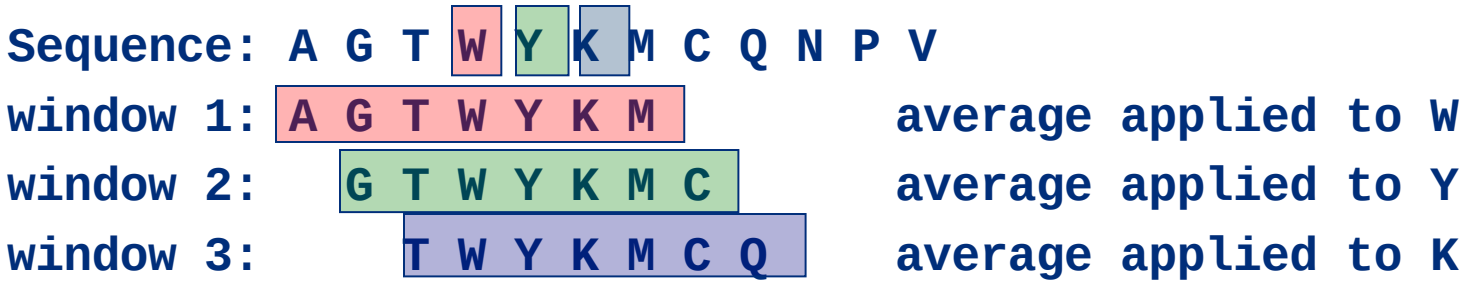
So, the helical propensity for subunit Alanine where:

- the number of alanines in the database is 1640,
- and the total number of subunits in the database is 10136,
- and where the number of alanines found in helices is 124,
- and the total number of subunits found in helices is 1246,

would be 0.61

Sliding windows

Propensity values are often assigned using sliding window methods



Theory that neighbouring subunits affect local structure

Example - Hydrophobicity

Hydro - phobic = water - hating

Some subunits do not exist happily in water - often on the inside of proteins

Some like water - take up positions on the outside of proteins.

This is also exploited in some structural elements such as helices

We can use a hydrophobicity propensity scale ...

Hydrophobicity Hash

```
%hydropathies = (
      A => 1.8, C => 2.5, D => -3.5, E => -3.5,
      F => 2.8, G => -0.4, H => -3.2, I => 4.5,
      K => -3.9, L => 3.8, M => 1.9, N => -3.5,
      P => -1.6, Q => -3.5, R => -4.5, S => -0.8,
      T => -0.7, V => 4.2, W => -0.9, Y => -1.3
      );
```

Each of the 20 protein subunits is assigned a value representing its hydrophobicity

Sliding window

```
@array = windowify($sequence);
```

```
sub windowify {
          @array = ();
          $startgap = int (7 / 2);
          $startpoint = 0;
          for ($h = $startgap; $h < ($seqlength - $startgap); $h++) {
                    $startpoint = $h - $startgap;
                    $array[$h] = calckds(substr($sequence, $startpoint, $window));
          }
          return @array;
};

sub calckds {
          $str = shift;
          @windowsection = unpack("A1" x length($str), $str);
          foreach $aa (@windowsection) {
                    $val += $hydropathies{$aa};
          }
          $val = ($val / $window);
          $val = int($val*1000)/1000;
          return $val;
};
```

Result

```
for ($i = 0; $i < $seqlength; $i++) {
        if ($array[$i] > 1.8) {
                $data_array[$i] = "-";
        } else {
                $data_array[$i] = "S";
        }
}
```

```
Sequence  MTTPTLIVTPPSPPAPSYSANRVPQPSLMDKIKKIAAIASLILIGTIGFLALLGHLVGFL
surface   SSSS------SSSSSSSSSSSSSSSSSS-S-------------------------

Sequence  IAPQITIVLLALFIISLAGNALYLQKTANLHLYQDLQREVGSLKEINFMLSVLQKEFLHL
surface   --------------------SSSS-SSSSSSSSSSSSSSSS----------S-SSSS

Sequence  SKEFATTSKDLSAVSQDFYSCLQGFRDNYKGFESLLDEYKNSTEEMRKLFSQEIIALKGS
surface   --SSSSSSSSSS--S--SS--SSSSSSSSSSSSSSSSSSSSSSSSSS---------

Sequence  VASLREEIRFLTPLAEEVRRLAHNQQSLTVVIEELKTIRDSLRDEIGQLSQLSKTLTSQI
surface   -SS-SSSSS---S-SSSSSSSSSSSS--------SSSSSSSSSSSSSSSS-SS-S---

Sequence  ALQRKESSDLCSQIRETLSSPRKSASPSTKSS
surface   SSSSSSSSSSSS-SSSSSSSSSSSSSSSSS
```

Cranfield UNIVERSITY

My Research

- Taken several propensity- style methods and applied them together
- Tailored analysis specifically for identifying target regions to bind antibodies
- Appear to be able to predict suitable regions >90% of the time

What now?

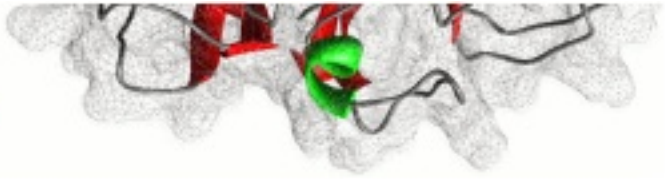
Analysed all 27,960 known human protein sequences - took 33 minutes (2Ghz MacBook)

Also several important bacterial species

Made a web- based tool and database for all this information.

AntigenFinder

the protein antigenicity repository



Main Menu
- Organism Databases
- Cancer Gene Search
- Custom Analysis
- About this project
- Home

Database Search

Search for proteins and sequences

This form can be used to search the database of protein sequences and the determined available binding regions. You can search either by name, genbank accession or short sequence. This is not a BLAST - it is literal sequence.

Enter a search term:

Database: Human

Name (•) Sequence ( )

Reset Submit

Main Menu
Organism Databases
Cancer Gene Search
Custom Analysis
About this project
Home

Database Record

Record for sequence: 15604838

Sequence Name
Inclusion Membrane Protein A

Species
Chlamydia trachomatis D/UW-3/CX

Genbank Identifier 15604838 to link to the genbank record click --> here
Genbank Accession NP_219622.1

Sequence Length: 273

Sequence

```
MTTPTLIVTPPSPPAPSYSANRVPQPSLMDKIKKIAAIASLILIGTIGFLALLGHLVGFL
IAPQITIVLLALFIISLAGNALYLQKTANLHLYQDLQREVGSLKEINFMLSVLQKEFLHL
SKEFATTSKDLSAVSQDFYSCLQGFRDNYKGFESLLDEYKNSTEEMRKLFSQEIIADLKG
SVASLREEIRFLTPLAEEVRRLAHNQQSLTVVIEELKTIRDSLRDEIGQLSQLSKTLTSQ
IALQRKESSDLCSQIRETLSSPRKSASPSTKSS
```

Below are the suitable antigens for this protein.

Sequence	Hydrophobics	Charged	Solubility	
CSQIRETLSSPRKSA	3	4	2	blast
DLCSQIRETLSSPRK	4	5	1	blast
ESSDLCSQIRETLSS	3	4	-2	blast
KESSDLCSQIRETLS	3	5	-1	blast
QKTANLHLYQDLQRE	4	5	1	blast
RETLSSPRKSASPST	3	4	2	blast
SQIRETLSSPRKSAS	3	4	2	blast