Biological Databases

Protein Structure and Function Analysis (SS 2011)

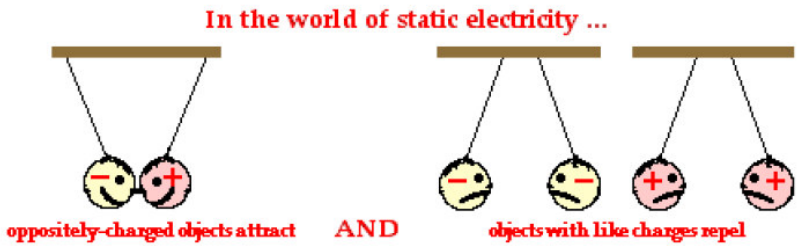
Contents

Contents

DSSP

- Assigns secondary structure to proteins
- No prediction!
- Requires 3D coordinates (from PDB files)

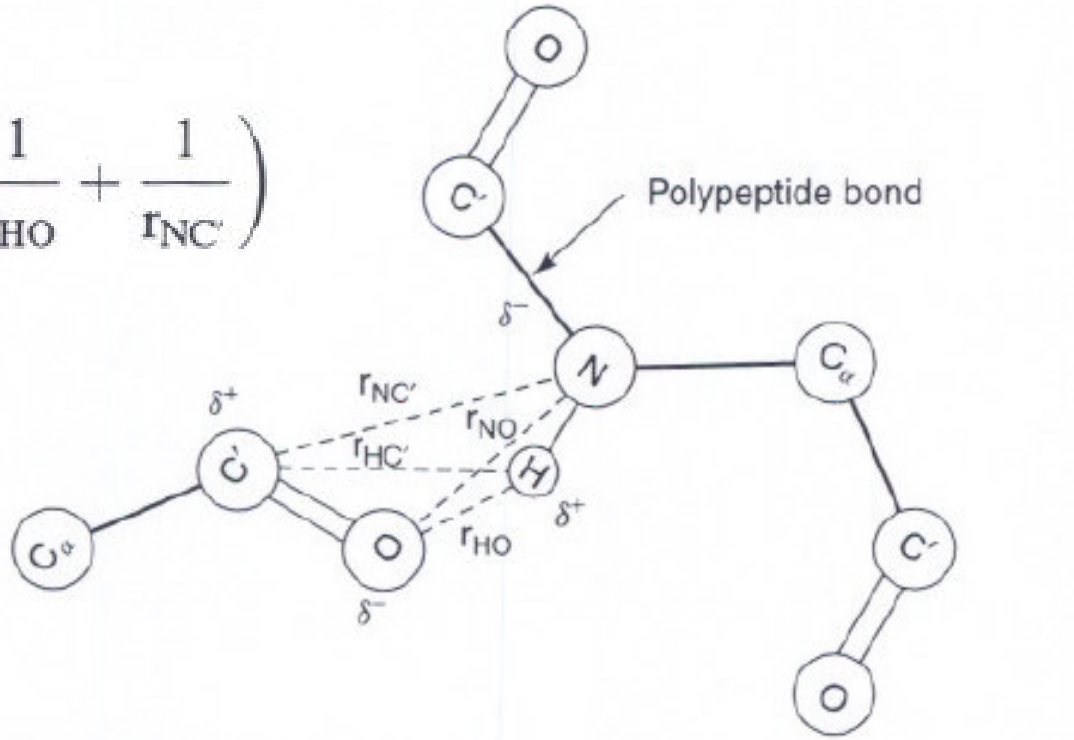
DSSP



- **Based on hydrogen binding calculations of backbone**
  - Employs coulomb's law
  - **f :** dimensional factor (332 A kcal/e²)
  - **δ+, δ- :** electron charges between N-C and O-H
  - **r:** distance

$$E = f\delta^{+}\delta^{-}\left(\frac{1}{r_{NO}} + \frac{1}{r_{HC'}} + \frac{1}{r_{HO}} + \frac{1}{r_{NC'}}\right)$$

Threshold: E < -0.5 kcal/mol

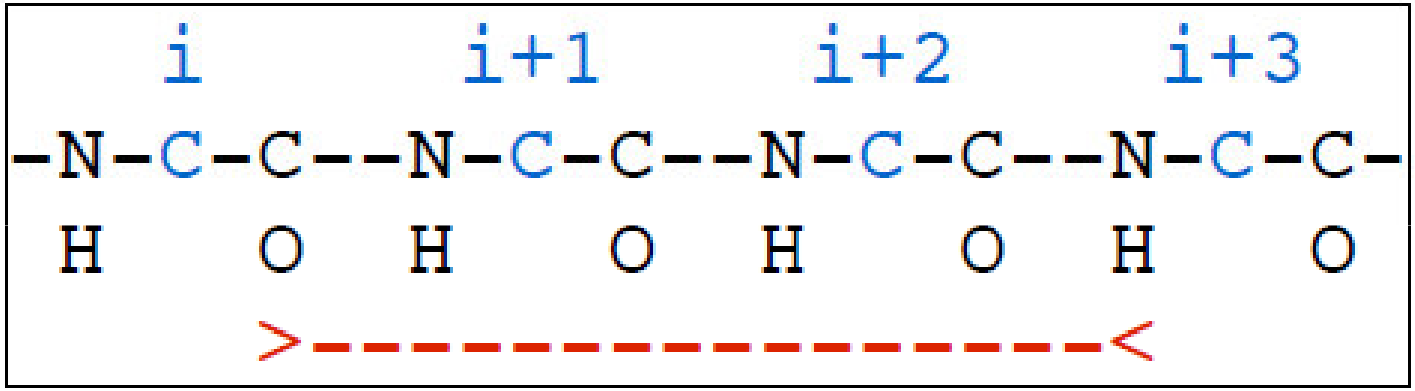


DSSP - Patterns

- SS assignment based on basic hydrogen binding patterns **„turn“** and **„bridge“**

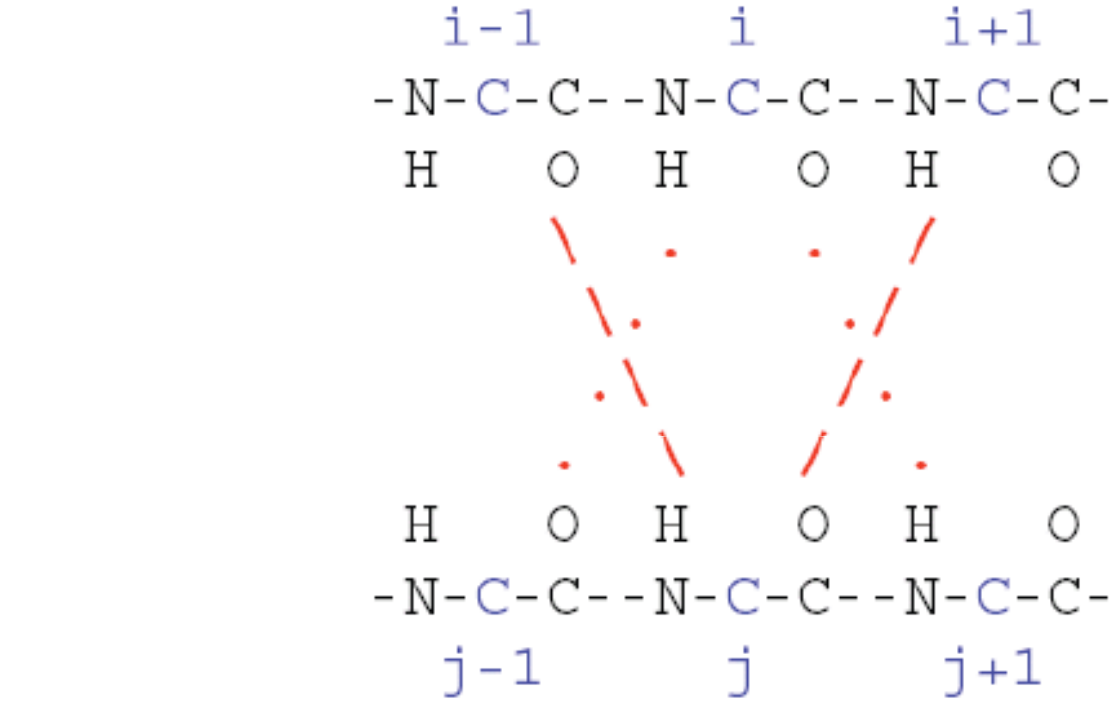
- **Multiple turns** = **„helices“**
- **Multiple bridges** = **„ladders“**
- **Interconnected ladders** = **„sheets“**

DSSP - Turns



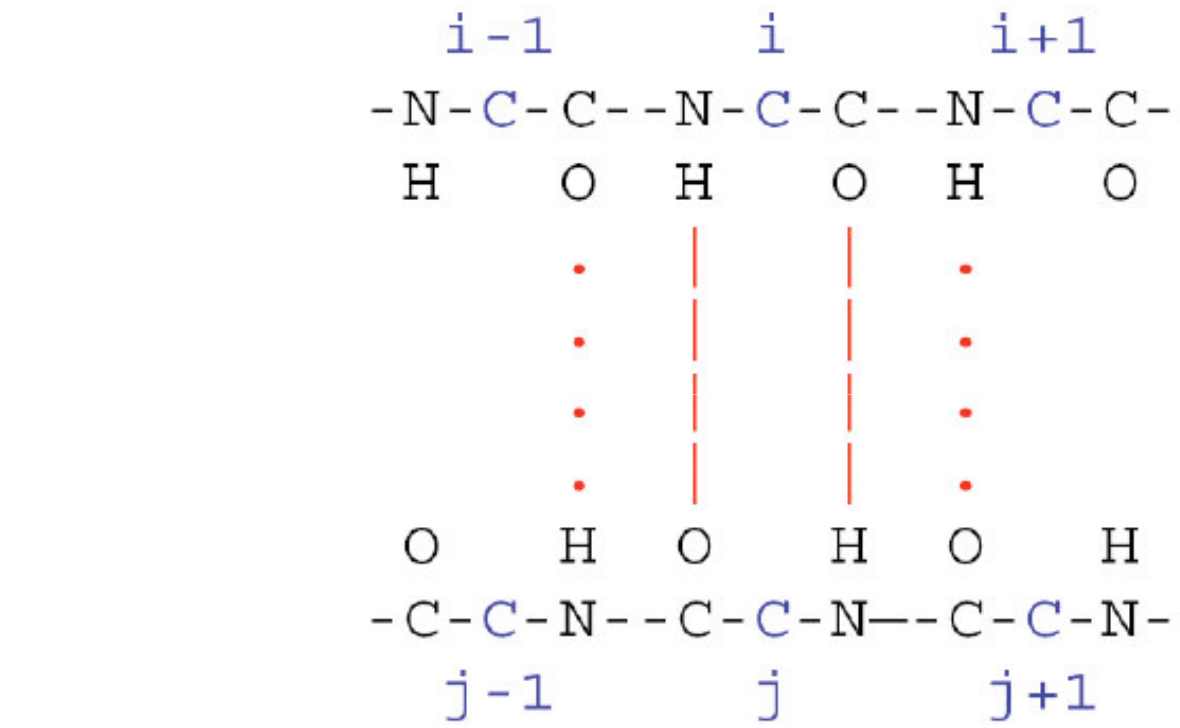
n-turn(i)=:Hbond(i,i+n), n=3,4,5

DSSP - Parallel Bridge



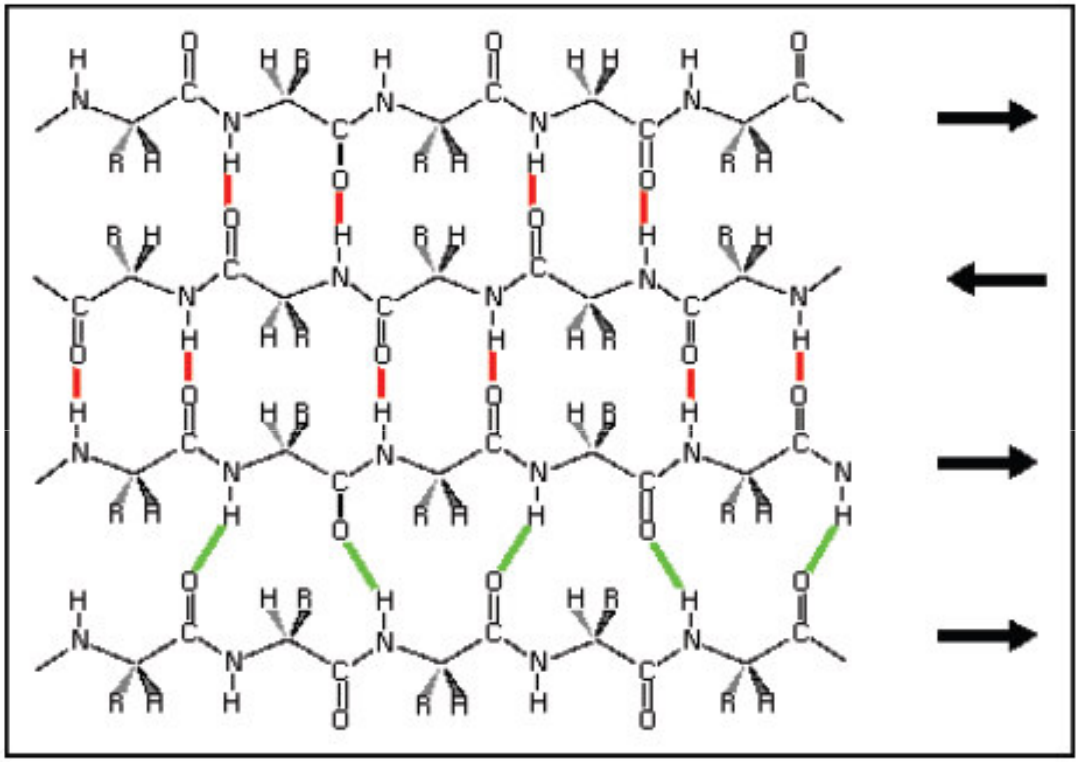
Parallel Bridge(i,j)=: [Hbond(i-1,j) and Hbond(j,i+1)] or
[Hbond(j-1,i) and Hbond(i,j+1)]

DSSP - Anti Parallel Bridge



Antiparallel Bridge(i,j)=: [Hbond(i,j) and Hbond(j,i)] or
[Hbond(i-1,j+1) and Hbond(j-1,i+1)]

DSSP - Ladders And Sheets



ladder=: set of one or more consecutive bridges
of identical type ,(anti-)parallel
sheet=: set of one or more ladders
connected by shared residues

DSSP – Residue Assignments

H = alpha-helix

B = residue in isolated beta-bridge

E = extended strand, participates in beta ladder

G = 3-helix (3/10 helix)

I = 5-helix (pi helix)

T = hydrogen bonded turn

S = bend

```
==== Secondary Structure Definition by the program DSSP, updated CMBI version by ElmK / April 1,2000 ==== DATE=20-MAR-2009
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
HEADER    OXIDOREDUCTASE                          22-MAY-94   1JIB
COMPND    2 MOLECULE: PHENYLALANINE-4-HYDROXYLASE;
SOURCE    2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
AUTHOR    O.A.ANDERSEN,T.FLATMARK,E.HOUGH
```

Basic Description

```
  307    1    0    0    0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
 14455.0    ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
  206 67.1    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
   15  4.9    TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
   14  4.6    TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    1  0.3    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
    2  0.7    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I),   SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
   12  3.9    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
   42 13.7    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
  110 35.8    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
    7  2.3    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
```

Statistical Metrics

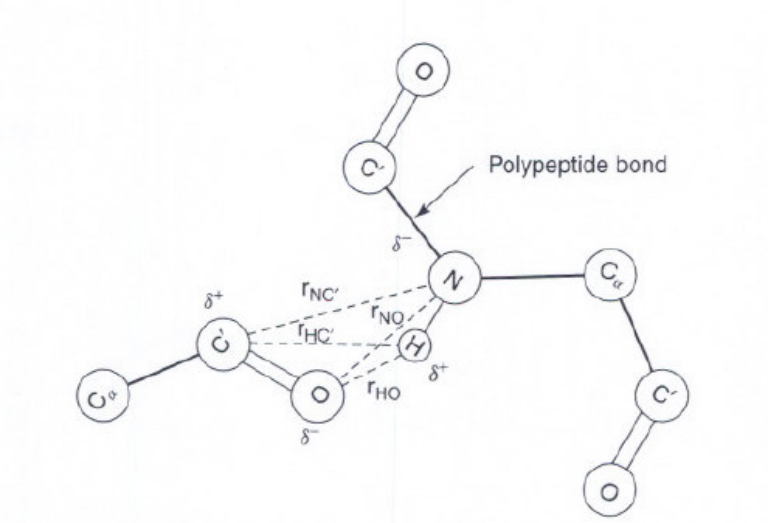
```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30      *** HISTOGRAMS OF ***
 0  0  0  3  0  0  2  0  1  0  0  2  1  2  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX
 1  1  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER
 2  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER
 1  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET
```

SS structure assignments for each residue

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
1	118 A	V		0	0	120	0, 0.0	195,-0.1	0, 0.0	3,-0.1	0.000	360.0	360.0	360.0	130.4	-22.9	34.1	18.4
2	119 A	P	-	0	0	48	0, 0.0	2,-0.0	0, 0.0	0, 0.0	-0.213	360.0	-100.8	-60.4	144.9	-21.3	31.1	16.7
3	120 A	W	+	0	0	105	191,-0.1	192,-0.3	192,-0.0	0, 0.0	-0.397	45.9	177.7	-63.0	144.5	-20.4	28.2	19.0
4	121 A	F	-	0	0	12	190,-0.1	2,-0.1	-3,-0.1	7,-0.0	-0.987	34.4	-97.0	-145.5	150.9	-16.7	28.0	19.9
5	122 A	P	-	0	0	5	0, 0.0	3,-0.1	0, 0.0	6,-0.0	-0.398	27.6	-175.3	-65.1	138.7	-14.7	25.6	22.1
6	123 A	R	S S+	0	0	120	1,-0.1	298,-2.6	-2,-0.1	2,-0.3	0.456	71.3	45.3	-113.1	-2.7	-14.0	26.9	25.6
7	124 A	T	B > S-A	303	0A	30	296,-0.2	3,-0.9	1,-0.1	4,-0.3	-0.953	84.8	-118.7	-132.5	156.2	-11.9	23.9	26.7
8	125 A	I	G > S+	0	0	0	294,-2.0	3,-1.5	-2,-0.3	4,-0.2	0.862	113.6	58.7	-64.0	-30.0	-9.0	22.3	24.9
9	126 A	Q	G > S+	0	0	94	1,-0.3	3,-2.2	293,-0.2	44,-0.3	0.823	90.2	70.8	-66.1	-29.6	-11.0	19.0	24.9
10	127 A	E	G X S+	0	0	67	-3,-0.9	3,-1.4	1,-0.3	-1,-0.3	0.574	75.4	81.9	-67.5	-9.3	-13.9	20.5	23.1
11	128 A	L	G X S+	0	0	0	-3,-1.5	3,-1.6	-4,-0.3	4,-0.4	0.705	73.5	79.5	-63.6	-17.6	-11.6	20.7	19.9
12	129 A	D	G X S+	0	0	43	-3,-2.2	3,-0.9	1,-0.3	4,-0.3	0.818	77.7	69.8	-57.0	-29.6	-12.7	17.0	19.6
13	130 A	R	G X S+	0	0	128	-3,-1.4	3,-1.0	1,-0.2	-1,-0.3	0.750	87.4	68.6	-58.1	-20.0	-16.0	18.3	18.1
14	131 A	F	G X S+	0	0	2	-3,-1.6	3,-2.1	1,-0.3	-1,-0.2	0.819	83.1	66.0	-77.0	-34.7	-14.1	19.4	15.0
15	132 A	A	G < S+	0	0	23	-3,-1.0	-1,-0.3	-4,-0.3	-2,-0.2	0.258	104.0	58.9	-71.2	-5.0	-13.3	16.1	13.5
16	133 A	N	G < S+	0	0	101	-3,-1.0	-1,-0.3	-4,-0.3	-2,-0.2	0.258	104.0	58.4	-106.1	25.0	-16.9	15.4	13.0
17	134 A	Q	S < S+	0	0	85	-3,-2.1	116,-0.5	2,-0.1	2,-0.2	-0.249	74.6	119.1	-131.0	41.0	-16.9	18.5	10.7
18	135 A	I	S S-	0	0	43	-3,-0.4	2,-0.7	114,-0.1	114,-0.2	-0.726	77.3	-91.5	-99.1	163.2	-14.3	18.0	7.9
19	136 A	L	-	0	0	93	112,-2.0	112,-0.3	-2,-0.2	3,-0.1	-0.613	41.5	-162.0	-63.6	111.7	-14.8	18.0	4.1
20	137 A	S	S S+	0	0	89	-2,-0.7	2,-0.7	1,-0.2	-1,-0.2	0.656	75.7	62.3	-76.9	-18.7	-15.4	14.2	3.7
21	138 A	Y	S S-	0	0	184	-3,-0.0	-1,-0.2	2,-0.0	-2,-0.0	-0.814	83.9	-171.7	-115.0	84.8	-14.8	14.1	-0.0
22	139 A	G	> -	0	0	15	-2,-0.7	3,-1.4	-3,-0.1	4,-0.2	0.120	38.4	-91.4	-81.6	-179.3	-11.2	15.2	-0.3
23	140 A	A	G > S+	0	0	31	1,-0.3	3,-1.7	2,-0.2	6,-0.1	0.782	114.9	69.6	-65.7	-26.9	-8.7	16.1	-2.9
24	141 A	E	G 3 S+	0	0	35	1,-0.3	-1,-0.3	4,-0.1	5,-0.1	0.628	92.5	61.0	-72.4	-12.4	-7.3	12.6	-3.2
25	142 A	L	G < S+	0	0	62	-3,-1.4	-1,-0.3	3,-0.0	2,-0.2	0.449	83.1	106.5	-86.2	-5.9	-10.7	11.6	-4.9
26	143 A	D	S X S-	0	0	81	-3,-1.7	3,-2.3	-4,-0.2	6,-0.2	-0.578	76.2	-127.7	-84.6	143.4	-10.2	14.0	-7.8
27	144 A	A	T 3 S+	0	0	84	1,-0.3	-1,-0.1	-2,-0.2	6,-0.1	0.778	108.6	53.0	-61.9	-27.7	-9.2	12.8	-11.2
28	145 A	D	T 3 S+	0	0	134	4,-0.1	-1,-0.3	-5,-0.1	-4,-0.1	0.415	79.3	118.4	-87.2	5.0	-6.3	15.1	-11.4
29	146 A	H	S X S-	0	0	9	-3,-2.3	3,-2.1	1,-0.2	4,-0.3	-0.539	72.6	-125.9	-72.0	139.1	-4.7	14.0	-8.1
30	147 A	P	T 3 S+	0	0	35	0, 0.0	3,-0.2	0, 0.0	-1,-0.2	0.791	111.4	40.0	-52.5	-29.8	-1.2	12.5	-8.6

DSSP – File Format

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
75	192	A	K H X S+	0	0	122	-4,-2.8	4,-1.1	-5,-0.2	-1,-0.2	0.904	115.4	48.0	-72.6	-36.5	17.1	17.4	0.6
76	193	A	T H < S+	0	0	45	-4,-1.8	3,-0.4	-5,-0.3	-2,-0.2	0.953	118.1	40.7	-69.5	-45.0	15.9	19.8	-2.1
77	194	A	L H >X S+	0	0	2	-4,-2.4	3,-2.7	-5,-0.3	4,-0.7	0.913	106.2	63.6	-67.0	-40.6	14.3	22.3	0.3
78	195	A	K H >X S+	0	0	80	-4,-2.4	3,-0.8	1,-0.3	4,-0.6	0.819	90.7	68.4	-57.1	-31.3	17.1	22.1	2.9
79	196	A	S H 3< S+	0	0	82	-4,-1.1	-1,-0.3	-3,-0.4	-2,-0.2	0.591	104.1	43.1	-59.4	-17.0	19.5	23.5	0.4
80	197	A	L H <> S+	0	0	10	-3,-2.7	4,-2.2	-4,-0.2	3,-0.4	0.575	90.9	82.4	-104.4	-15.7	17.7	26.9	0.6
81	198	A	Y H <X S+	0	0	6	-3,-0.8	4,-2.3	-4,-0.7	9,-0.2	0.898	82.1	57.8	-69.1	-39.0	17.0	27.4	4.2
82	199	A	K H < S+	0	0	184	-4,-0.6	-1,-0.2	1,-0.2	-2,-0.1	0.870	121.8	29.1	-56.2	-37.2	20.3	28.9	5.5



DSSP - Accession

- **FTP**

```
$ wget ftp://ftp.cmbi.ru.nl//pub/molbio/data/dssp/*
```

```
$ wget ftp://ftp.cmbi.ru.nl//pub/molbio/data/dssp/<PDB ID>.dssp
```

- **Rsync**

```
$ rsync -avz rsync://rsync.cmbi.ru.nl/dssp/ .
```

DSSP - Summary

- **What is DSSP?** DB of secondary structure **assignment** of proteins with solved 3D structure. **No prediction**!
- **Why do we need DSSP?** For visualisation of proteins, multiple sequence alignments, topology of proteins and many more ...
- **What assumption is exploited?** Based on patterns of simple elements like “turn” and “bridge” which are derived from hydrogen bond calculations.
- **How can I use it?**
  - **Listening of all entries:**

```
$ wget ftp://ftp.cmbi.ru.nl//pub/molbio/data/dssp/*
```

  - **Information of one PDB structure:**

```
$ wget ftp://ftp.cmbi.ru.nl//pub/molbio/data/dssp/<PDB ID>.dssp
```

Contents

1. DSSP
2. HSSP
3. UniProt

HSSP – DSSP's big brother

- **What is HSSP?** DB of merged information from three-dimensional structures and one-dimensional sequences of proteins
- **Why do we need HSSP?** If only one 3D structure of one family member is known the 3D structure/fold of all close family members can be derived
- **What assumption is exploited?** Stems from the evolutionary observation that protein sequences can vary considerably while maintaining the same overall 3-D structure
- **How can I use it?**
  - **Listening of all entries:**

```
$ wget ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/hssp
```

  - **Information of one PDB structure:**

```
$ wget ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/hssp/<ID>.hssp
```

Contents

1. DSSP
2. HSSP
3. UniProt

UniProt

- **UniProtKB**
  - UniProtKB/SwissProt: manually annotated
  - UniProtKB/TrEMBL: High quality computational annotation
- **UniRef:** sequence clusters of 100, 90, 50%
- **UniParc:** contains all new and revised protein sequences from all publicly available sources.
- **UniMES:** Metagenomic and Enviromental Sequence Database

UniProt – The Bioinformaticians Way

- **FTP:** ftp.uniprot.org/pub/databases
  - Download of all entries

- **REST:** a stateless way of querying UniProt via a URL
  - http://purl.uniprot.org/uniprot/P12345.txt
    - get the Entry P12345 in text file format
    - Possible Formats: TXT, XML, RDF, FASTA, GFF

UniProt – The Bioinformaticians way cont'd.

- Retrieve all reviewed entries from human:

```
$ wget http://www.uniprot.org/uniprot/?query=reviewed:yes+AND+organism:9606&format=xml
```

- Possible parameters:

Query: string
Format: html | tab | xls | fasta | gff | txt | xml | rdf | list | rss
Columns: citation | clusters | comments | database | domains | …
Compress: yes | no
Limit: integer
Offset: integer

UniProt – The Bioinformaticians way cont'd.

- **ID Mapping in Python:** also for Perl, Ruby and Java

```
import urllib,urllib2

url = 'http://www.uniprot.org/mapping/'
params = {
       'from':'ACC',
       'to':'P_REFSEQ_AC',
       'format':'tab',
       'query':'P13368 P20806 Q9UM73 P97793 Q17192' }

data = urllib.urlencode(params)
request = urllib2.Request(url, data)
response = urllib2.urlopen(request)
page = response.read(200000)
```

UniProt – The Bioinformaticians way cont'd.

- **UniProtJAPI**: Java API to access UniProt
  - **blast/search against UniProtKB, UniParc, UniRef or UniMes possible**

- **UniProt BioMart:** querying UniProtKB data, with a cross-querying facility to join to/from other BioMart databases (currently Ensembl Genes and the PRIDE Proteomics Identifications Database).

- **Swissknife:** Object-oriented Perl library for parsing the UniProtKB text format.

Resources/References

- **DSSP**
  - Website: http://swift.cmbi.ru.nl/gv/dssp/
  - Detailed explanation of output: http://swift.cmbi.ru.nl/gv/dssp/
  - DSSP Publication: http://swift.cmbi.ru.nl/gv/dssp/HTML/dssp.pdf
  - Explanation on Coulomb Hydrogen Bond Calculations: http://www.inanoschool.au.dk/fileadmin/inano/iNANOSchool/iNANOschool_2006/N9/structural_bioinformatics.pdf
  - Coulomb‘s law: http://en.wikipedia.org/wiki/Coulomb%27s_law
  - Lecture materials of „Structural Bioinformatics“ held by Prof. Frishman

- **HSSP**
  - Website: http://swift.cmbi.kun.nl/swift/hssp/
  - Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/9399862

Resources/References – cont‘d

- **UniProt**
  - Website: http://www.uniprot.org/
  - Pubmed: http://www.ncbi.nlm.nih.gov/pubmed/19843607
  - Technical Information on UniProt: http://www.uniprot.org/help/technical
  - Detailed description of query fields: http://www.uniprot.org/help/query-fields
  - REST access: http://www.uniprot.org/faq/28
  - UniProtJAPI: http://www.ebi.ac.uk/uniprot/remotingAPI/
  - Swissknife: http://swissknife.sourceforge.net/docs/