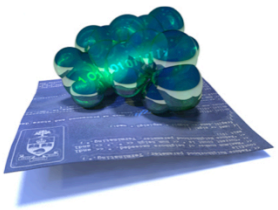
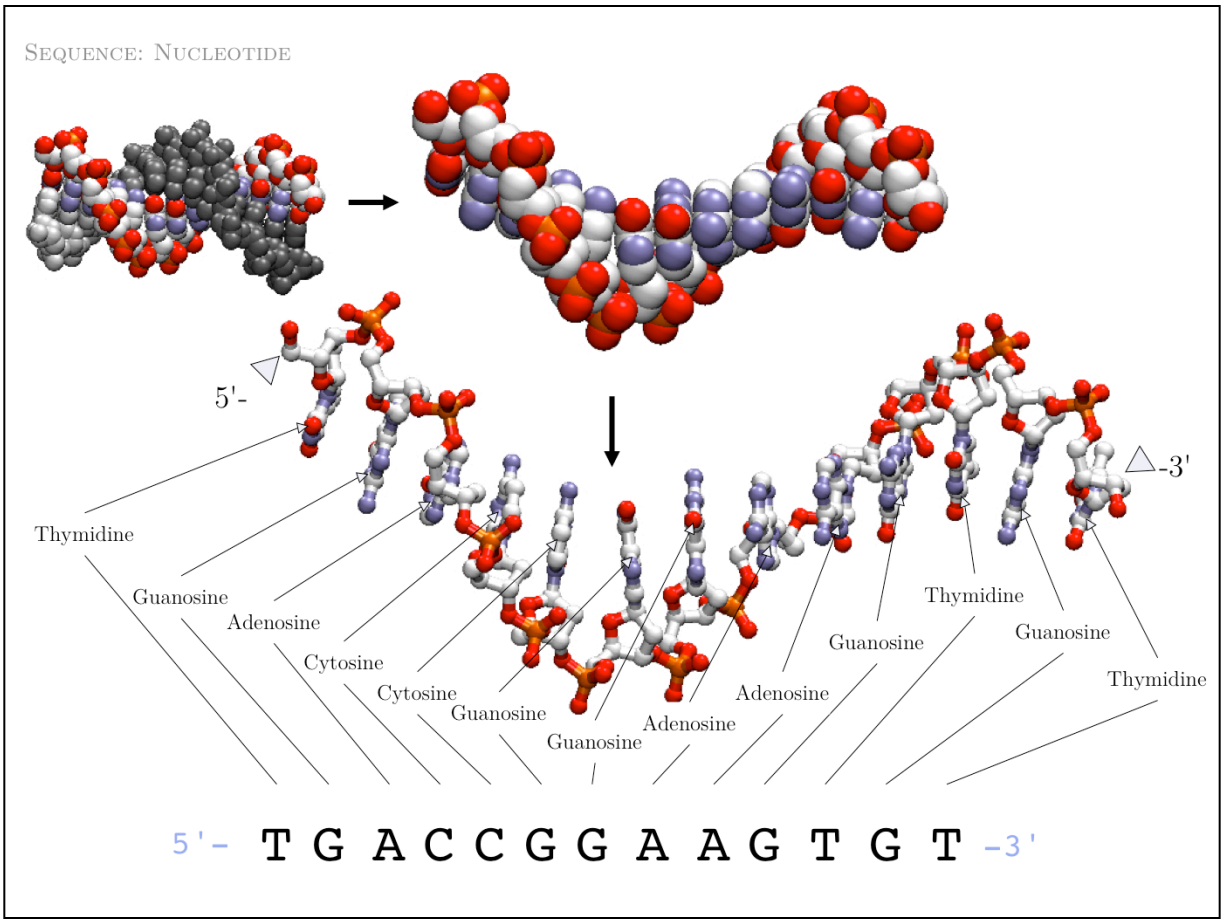
A
Bioinformatics
Course

Sequence



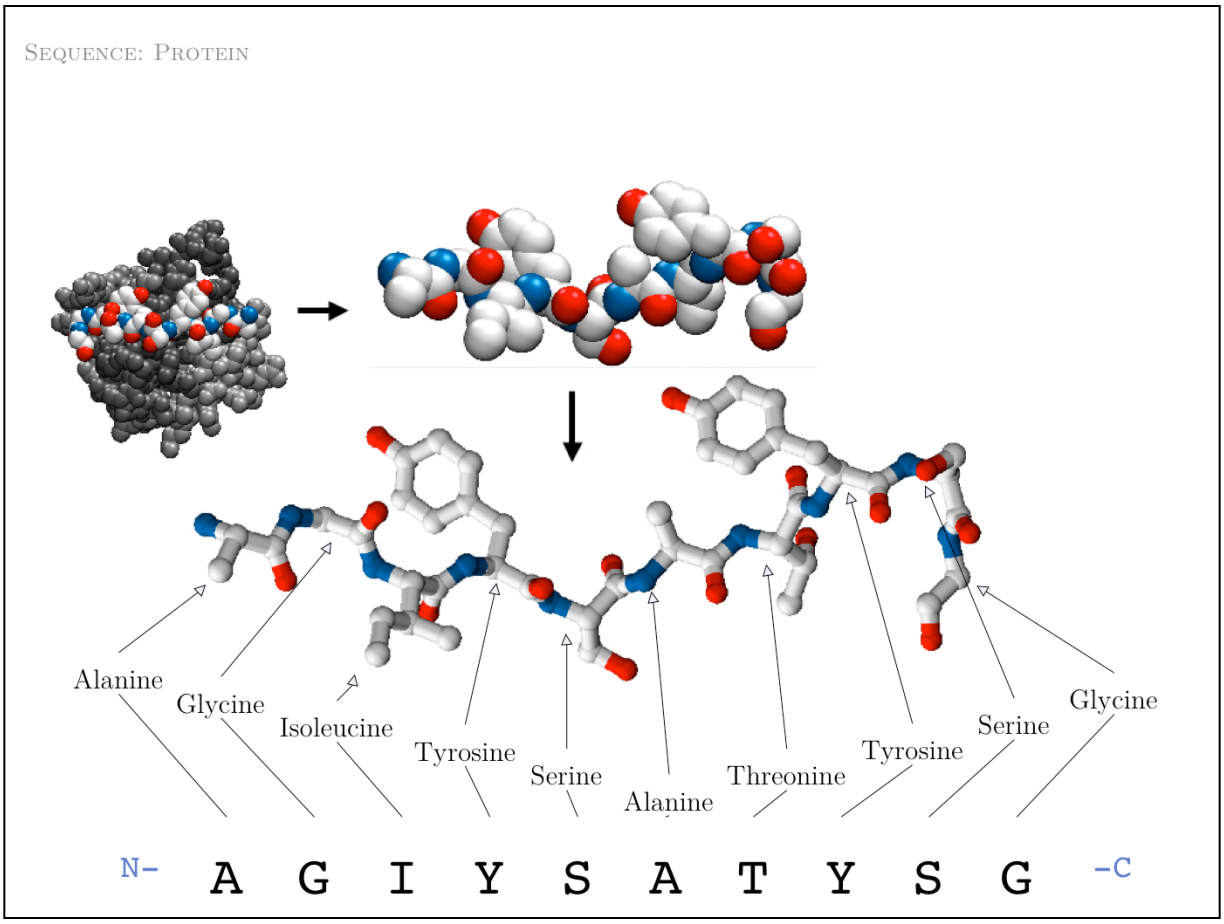
Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



Nucleic acids can form heterocopolymers as DNA or RNA, thus their structural formula can be described (to a close approximation) simply by listing the nucleotide bases in a defined order. A one-letter code has been defined as a shorthand notation for this. By convention, DNA or RNA sequence is written in the direction from the bond with the 5' carbon of the ribose (or deoxyribose), to the bond with the 3'-carbon (identified by grey triangles in the image). Since the 5'-carbon carries a free 5'-phosphate at the terminus of the polynucleotide and the 3'-carbon carries a free OH, this direction happens to be the same direction that nucleic acid polymers are replicated by the DNA-polymerase: a phosphate of a single nucleotide is attached to the free -OH of the polynucleotide. This direction is also the direction of transcription of the RNA polymerase (for the same reason) and (incidentally) the direction of translation by the ribosome (that would not **have** to be the case, but it allows immediate, co-transcriptional translation). Due to base-pair complementarity, only one strand of a double stranded sequence needs to be recorded since the sequence of the complementary strand is implied: it is simply the reverse complement. The one letter abbreviations are defined by the International Union of Pure and Applied Chemistry (IUPAC) as follows:

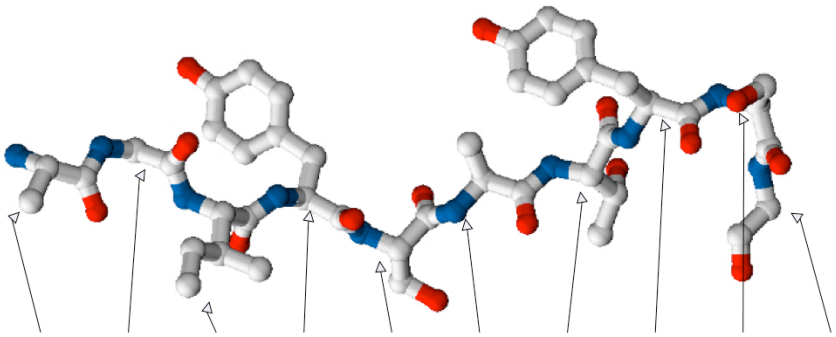
A = Adenine	R = G A (puRine)	B = G T C (B is not A)
C = Cytosine	Y = T C (pYrimidine)	D = G A T (D is not C)
G = Guanine	K = G T (Keto)	H = A C T (H is not G)
T = Thymine	M = A C (aMino)	V = G C A (V is not T)
	S = G C (Strong bonds)	
	W = A T (Weak bonds)	N = A G C T (aNy)



Proteins are amino acid heterocopolymers, thus their structural formula can be described (to a close approximation) simply by listing their constituent amino acid residues **in a defined order**. In fact, we believe that **all** of the properties of a protein are defined by its sequence (but we quietly acknowledge to ourselves that there are exceptions to this rule in biology, as there are to any rule in biology).

The one-letter code is a shorthand notation for the structural formula of a protein. By convention, protein sequence is written from amino terminus (N-) to carboxy terminus (-C). This happens to be the same direction in which we write the coding DNA sequence, and the same direction in which the protein is synthesized on the ribosome.

Sequence



A G I Y S A T Y S G

A sequence has: composition, length, direction, order ...

A sequence implies: a gene, or a fragment thereof, a protein, a structure, a function ...

Definition:
A sequence is an ordered set of letters that represents the structural formula of a biological polymer by listing its building blocks.

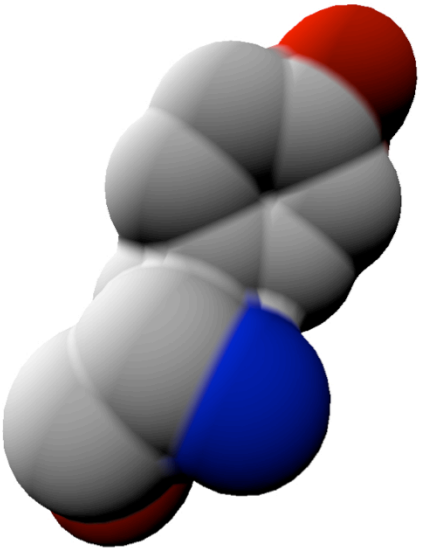
The most common abstration for protein sequence is a string of characters of the 1-letter amino acid code, ordered from N-terminus to C-terminus.

> Sequence models nature's own "abstraction" of linear, ordered arrangements of codons to store information; it is straightforward to phrase a model of evolution in terms of sequence change.
>
> Sequences are compact, easy to understand and contain (in principle) sufficient information to specify a protein's structure and function.
>
> Sequences are easy to compute with; their manipulation, comparison and algorithmic analysis is well understood.

Since sequences are so ubiquitous in the life sciences, you must be able to read them, *i.e.* to recognize the relevant amino acid properties for each character.

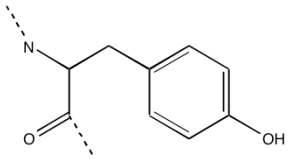
If you don't know the 1-letter code by heart, you may be doing informatics with sequences, but you probably won't be doing **bio**informatics.

AMINO ACID ABSTRACTION

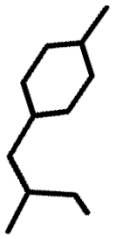


Formula: $C_9H_9NO_2$

Smiles String: [CH]([NH][R])([C](=[O])[R])[CH2]-[c]1([cH][cH][c]([cH][cH]1)[OH])



Name: Tyrosine



3-Letter: Tyr

1-Letter: Y

```
ATOM   1091  N   TYR   145     -35.676 -13.136  50.622  1.00 10.36
ATOM   1092  CA  TYR   145     -36.931 -13.763  51.019  1.00 10.63
ATOM   1093  C   TYR   145     -37.676 -12.879  52.016  1.00 11.16
ATOM   1094  O   TYR   145     -37.061 -12.316  52.926  1.00 13.91
ATOM   1095  CB  TYR   145     -36.660 -15.140  51.638  1.00  9.52
ATOM   1096  CG  TYR   145     -37.845 -15.737  52.361  1.00  6.36
ATOM   1097  CD1 TYR   145     -38.144 -15.357  53.663  1.00  3.30
ATOM   1098  CD2 TYR   145     -38.691 -16.652  51.727  1.00  6.14
ATOM   1099  CE1 TYR   145     -39.248 -15.856  54.311  1.00  5.57
ATOM   1100  CE2 TYR   145     -39.804 -17.165  52.376  1.00  4.89
ATOM   1101  CZ  TYR   145     -40.076 -16.757  53.670  1.00  4.35
ATOM   1102  OH  TYR   145     -41.170 -17.231  54.345  1.00  4.44
```

An amino acid is a molecule. A number of abstractions are in common use for such molecules:

The chemical formula simply describes the elemental composition.

The so called SMILES string[1] captures bonding topology as well; its information is equivalent to a chemical graph.

A set of records of 3D coordinates can describe the three-dimensional conformation, this can in turn be displayed in a number of different image options - like a simple line drawing or a set of spheres, color coded by element, with relative sizes corresponding to the elements' Van der Waals radii.

The simplest abstraction is the 1-letter code.

A useful compilation of amino acid properties can be found on Wikipedia[2].

[1] http://en.wikipedia.org/wiki/Simplified_Molecular_Input_Line_Entry_Specification

1-LETTER CODE

Amino acid structure is implied by the 1-letter code :

...plus the following ambiguity codes:

B: D/N

J: I/L

O: pyrrolysine

U: selenocysteine

X: unknown

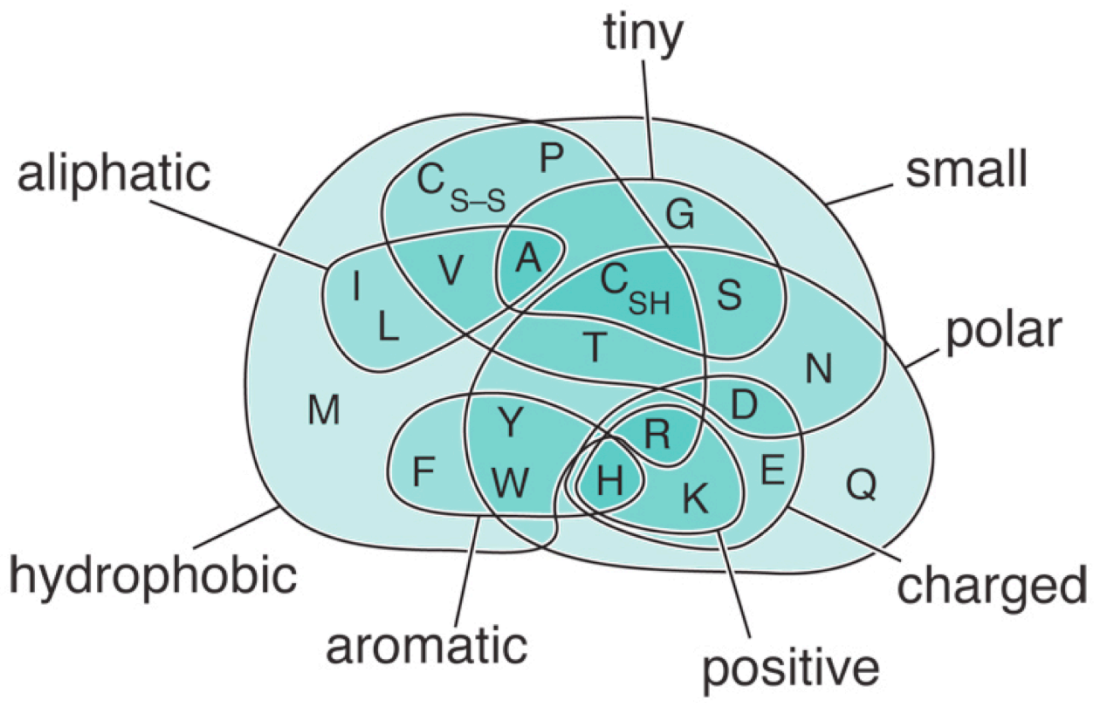
Z: E/Q

Sequence is the most important abstraction in biology; you need to know your amino acids in order to relate a sequence back to the biopolymer. **Required knowledge is: the structural formula, the one- and three- letter codes** and **key properties (such as charge, relative size, polarity) for all 20 proteinogenic amino acids.** You must commit this to memory.

A resource that summarizes amino acid properties is at `http://en.wikipedia.org/wiki/Amino_acid`

Similarity

Biophysical properties provide a first-order approach to define amino acid similarity.



Obviously, the precise role of a particular amino acid depends on its context in a folded protein, however this Venn diagram (originally going back to Willie Taylor) provides a good first aproximation to summarize shared sidechain properties and to estimate amino acid similarity.

Note that "C" appears twice in this sketch: once as cysteine (C_{SH}) with its free thiol function, once as the disulfide bonded cystine ($C_{S\text{-}S}$). These two forms have very different properties.

Consider the following sequence:

... AGSWCGGCRYGA...

What does it signify?

Is this to be read as peptide? As DNA? As RNA?

Some ambiguity can only be resolved if the context of the representation is specified. If these are amino acids, this can be unambiguously tranlated into a peptide sequence. However these could also be IUPAC nucleotide (ambiguity) codes.

Information can only be retrieved from computer representations if the semantics of the encoding are defined!

Abstracting biomolecules as sequences is fundamental to bioinformatics – but there are important limitations to be aware of:

- Protein/DNA/RNA sequence can be *ambiguous*.
- The limited standard alphabet cannot represent *modifications*.
- There is no implied, stable system for defining *sequence numbers* so they can be transferred between different sequences.
- Sequences alone cannot store *annotations*.

An ordered set of letters to represent biomolecules has some obvious limitations.

The issue of sequence numbering is especially odious: not only can the starting residue vary between closely related sequences, related sequences often have insertions or deletions relative to each other. Thus, if a journal article mentions eg. a `K25` residue, it may not be entirely clear which residue exactly this refers to. Some communities use canonical sequence numbers from reference domains, and use insert codes or skip numbers where necessary (e.g. immunoglobulins).

Therfore: even though a sequence number is unambiguously defined with respect to a sequence, it may not be particularly useful, and even ambiguous in practice. These however break the assumption that such "numbers" are sequential integers. Minimally, when specifying a “sequence number”, you need to also specifiy the precise sequence that the number refers to.

Y

What does this mean?

Tyrosine?

Pyrimidine?

Yes?

Why?

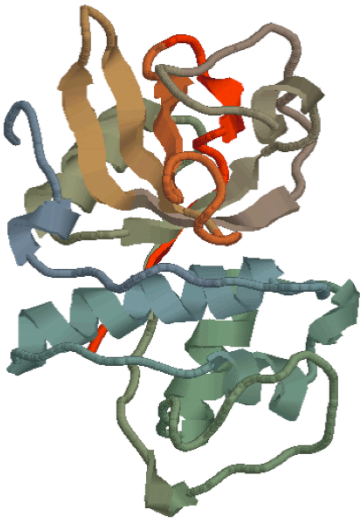
Letters have a meaning in bioinformatics **only** if we specifiy the semantics of what they represent.

Conventions to encode sequence

- IUPAC nucleotide code;
- IUPAC amino acid code;
- 5'- to 3'-, or N- to C-;
- sense- antisense strand vs. plus/minus, top/bottom (Watson/Crick);
- For genomes, only the plus strand is stored in the database.
- The A allele of a SNP is on the "plus" strand;
- Chromosome regions;
- units: BP, KB, MB, GB (kilo- mega- giga- base pairs).z

- DNA/RNA sequences are encoded in the IUPAC one letter nucleotide code.
- Protein sequences are encoded in the IUPAC one letter amino acid code that you need to memorize.
- The sense or + strand is the strand that could be translated. The complementary strand is called antisense, or "-" strand. This has nothing to do with which strand is written on the top or bottom if both strands are shown, although conventionally the sense strand is written on top. Sometimes you may find the labels "Watson" (top strand) and "Crick" strand (bottom strand). This is obfuscating nonsense, but appears as "W" or "C" letters at the end of yeast systematic gene names. Thus "forward", "top", "plus" and "Watson strand" are synonyms;
- For genomes, only one strand, the **plus** strand is stored in the database – but **which** strand is defined to be the plus strand is arbitrary.However, human chromosomal coordinates start at the telomere of the short arm, and extend to the telomere of the long arm.
- Chromosome regions are assigned the letter p or q, depending on whether they are on short or long arm of the chromosome; Giemsa stain banding number from the centromere outward further resolve the region. For example "7p14.1" is chromosome 7, short arm, band 14 from the telomere out, subregion 1 and this regio is defined to cover coordinates 37,100,001 to 43,300,000.

Example:
Cathepsin K
Cathepsin O, X, O2
Lysosomal, secreted cysteine protease
Tissue remodeling - role in ECM degradation and bone desorption
Pre(15aa)-Pro(90aa)-Protein(215aa)

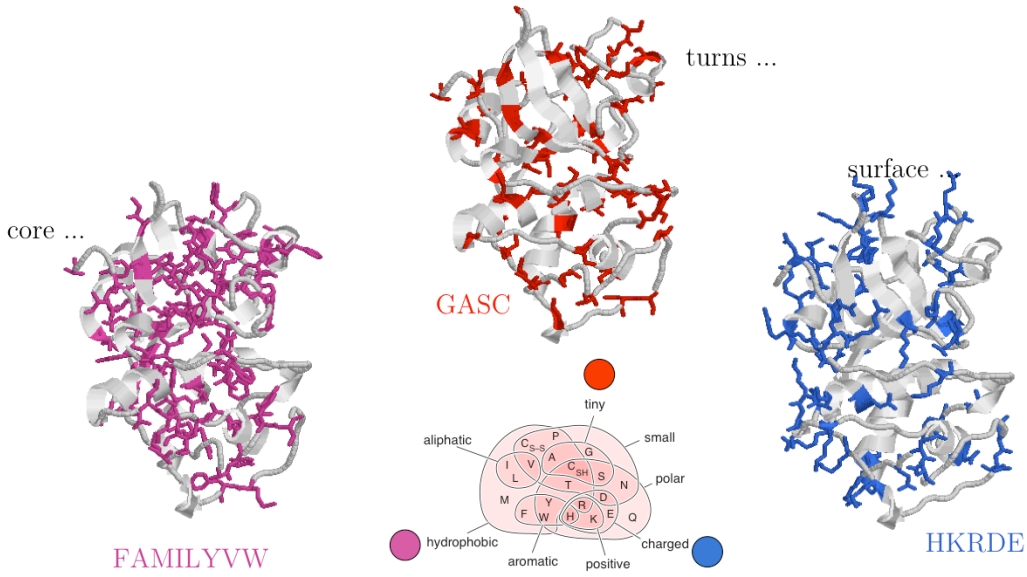


1ATK.PDB

The physicochemical properties of amino acids determines their role in e.g. a folded protein structure. According to their structural and functional roles, different amino acids will be found in different parts of the structure

For example, consider the amino acid distribution in a typical enzyme, such as cathepsin K (1ATK)...

Amino acids are non-randomly distributed in protein structure



Similar **properties** of amino acids lead to similar **distribution** in protein structure.

- Hydrophobic amino acids - the group `FAMILYVW` - are found predominantly in the core of a protein.
- Small amino acids such as `GASC` are often found in turns, at the boundaries of secondary structure elements
- charged amino acid sidechains – (+)`KRH` and (-)`DE` – are almost exclusively found on the surface; the energetic requirements for desolvation of the sidechain makes their incorporation into the core unfavourable.

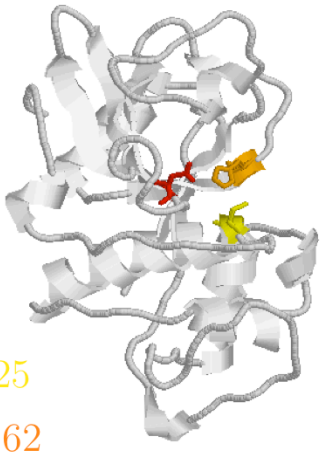
Active site cysteine

25

```
CATK_HUMAN ...  GSCWAF ...
PAP3_CARPA ...  GSCWAF ...
ORYB_ORYSA ...  GSCWAF ...
CYSL_LYCES ...  GSCWAF ...
CYSP_HEMSP ...  GSCWAF ...
CATL_DROME ...  GSCWAF ...
CATJ_RAT   ...  GSCWAF ...
ALEU_HORVU ...  GSCWTF ...
BROM_ANACO ...  GACWAF ...
EUM1_EURMA ...  GSCWAF ...
CPR5_CAEEL ...  GSCWAF ...
CYS1_OSTOS ...  GSCWAV ...
```

Related protein sequences

http://pfam.wustl.edu/



Cys25

His162

Asn182

1ATK.PDB

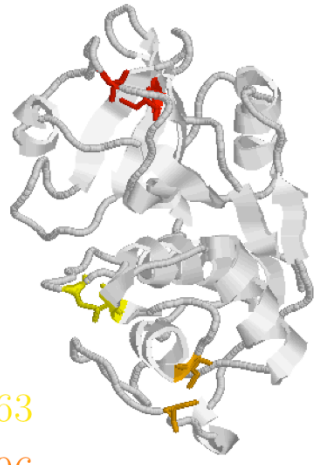
Cysteine can take on a number of different roles, depending on its context. In cathepsin, cysteine forms part of the active site, it is the nucleophile in the catalytic triad C-H-N; cathepsin is thus an example of a cysteine-protease. This particular cysteine is **absolutely conserved** in related proteins since its substitution would lead to (nearly) complete loss of enzymatic function.

Disulfide bonded cysteine

```
                          56          63          96
                          |           |           |
CATK_HUMAN  ... NLVDCVSE...ND.GCGGGY  ... SCM ...
PAP3_CARPA  ... ELVDCER...RSH.GCKGGY  ... TCR ...
ORYB_ORYSA  ... ELVECSTNG.QNS.GCNGGL  ... KCD ...
CYSL_LYCES  ... ELVDCDR.S.YNE.GCDGGL  ... VCD ...
CYSP_HEMSP  ... ELVDCDKEE..NQ.GCNGGL  ... TCD ...
CATL_DROME  ... NLVDCST.KYGNN.GCNGGL  ... SCH ...
CATJ_RAT    ... NLLDTKSE...GI.GLPWGT  ... PCR ...
ALEU_HORVU  ... QLVDCAG.GFNNF.GCNGGL  ... VCH ...
BROM_ANACO  ... QVLDCAK....GY.GCKGGW  ... TCK ...
EUM1_EURMA  ... ELVDCAS....QN.GCHGDT  ... SCH ...
CPR5_CAEEL  ... DLLSCCTGMFSCGNGCEGGY  ... KCV ...
CYS1_OSTOS  ... DVVSCCTWCGD...GCEGGW  ... RCK ...
```

Related protein sequences:

http://pfam.wustl.edu/



22–63

56–96

155–204

1ATK.PDB

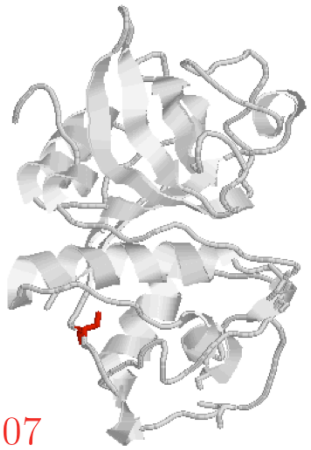
In secreted proteins only, cysteine often forms structural disulfide bridges in which two thiol groups oxidize to a covalent disulfide bond. These cysteines usually are highly conserved.

Proteins that are localized in the reducing environment of the cytoplasm do not form structural disulfide bridges.

"General" cysteine

```
                  107
                  |
CATK_HUMAN ... AKCRGY
PAP3_CARPA ... VKTSGV
ORYB_ORYSA ... VSIDGF
CYSL_LYCES ... VKIDSY
CYSP_HEMSP ... VSIDGH
CATL_DROME ... ATDRGF
CATJ_RAT   ... ANITGF
ALEU_HORVU ... VQVLDS
BROM_ANACO ... Y.ITGY
EUM1_EURMA ... YGLKNY
CPR5_CAEEL ... LQDKHF
CYS1_OSTOS ... PSDRYY
```

Related protein sequences:
http://pfam.wustl.edu/



Cys107

1ATK.PDB

Cysteine can also be found in a very general role, simply as a somewhat polar, small residue. Cysteines in such a general role are only seen infrequently in secreted proteins since the unpaired cysteines can interfere with the formation of the correct disulfide topology; this can lead to slow folding and generally makes the protein sensitive to oxidation. Such cysteines are poorly conserved in related proteins.

Despite all its limitations, the abstraction "sequence" is by far the most useful and widely used representation of biopolymers:

Sequence models nature's own "abstraction" of linear, ordered arrangements of codons to store information; it is straightforward to phrase a model of evolution in terms of sequence change.

Sequences are compact, easy to understand and contain (in principle) sufficient information to specify a protein's structure and function.

Sequences are easy to compute with; they are discrete entities and their management, manipulation and comparison are algorithmically well understood.

Sequence Change

Genome

- Point mutation (silent, missense, nonsense)
- Insertion and deletion
- Translocation
- Gene fusion
- Inversion
- Duplication
- WGD
- Horizontal Gene Transfer

Transcriptome

- Alternate transcriptional start and stop
- RNA editing
- Splicing, alternate splicing

Proteome

- Alternate translation
- Post-translational modification
- Peptide fusion
- Signal peptide cleavage
- Proteolysis

Sequences in biology are not static and a large number of processes act to modify sequences in the course of evolution as well as during the normal function of the cell. Some of these processes generate problems for representing sequences; for example sequence numbers may depend on alternate splicing, or cleavage of pre- or pro-sequences. As well, most of the (very common!) post-translational modifications cannot be mapped to the 20-letter code. You should be familiar with these processes.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA