Advanced Bioinformatics

Development of algorithms
Genomic analysis

Machine learning techniques
Features, applications and validation procedures

Research themes

Classification problems

Helix, strand and loop (secondary structures)

DNA/RNA binding proteins or NOT

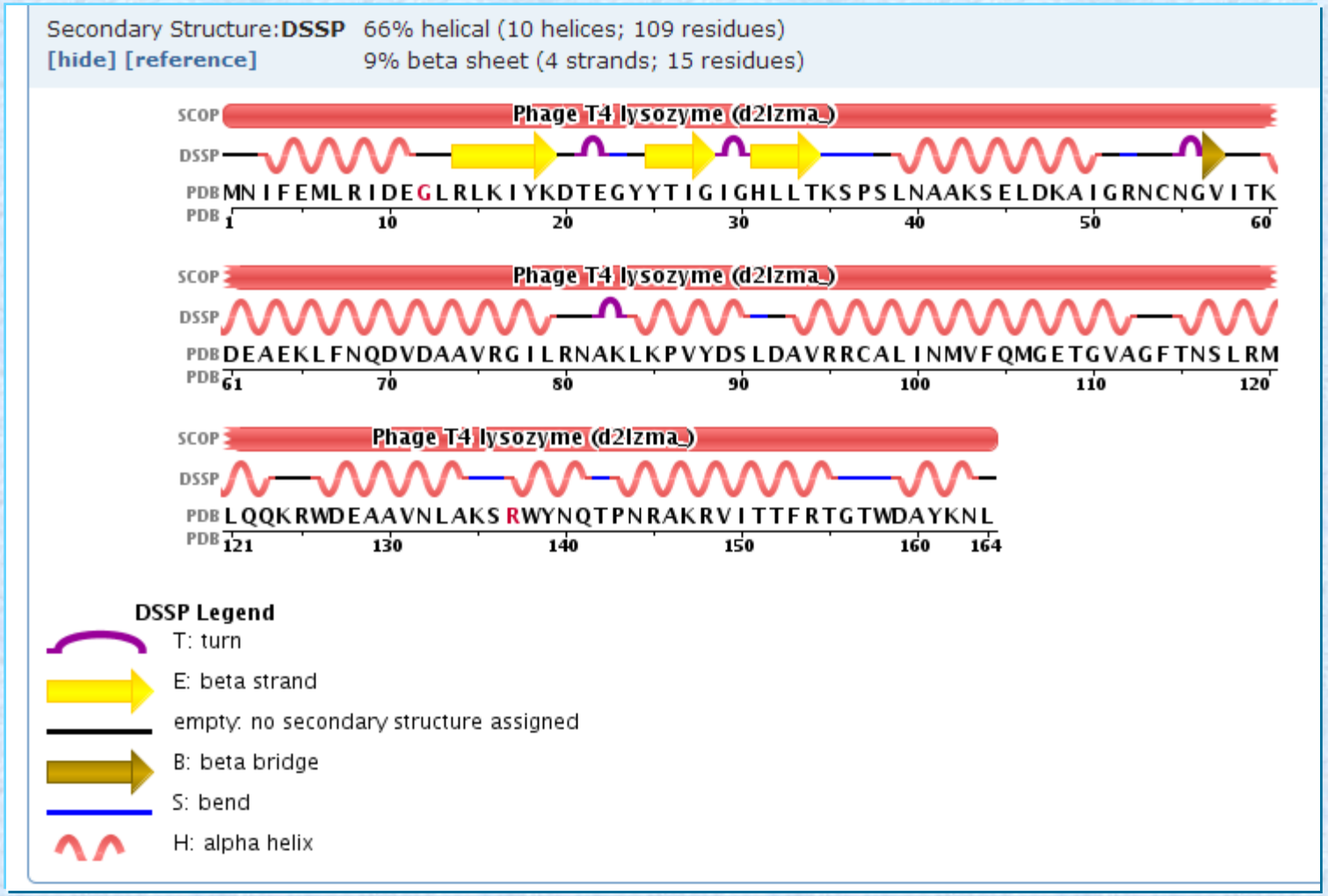
Binding sites or NOT

Residues in membrane or NOT

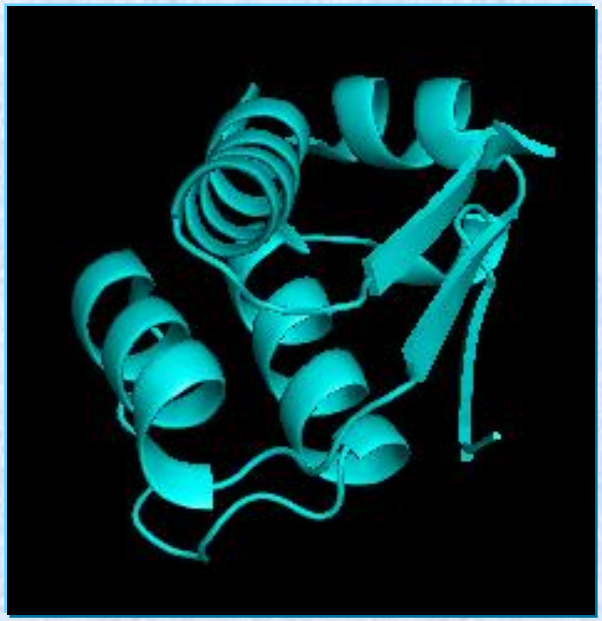
Perform specific function (transporters) or NOT

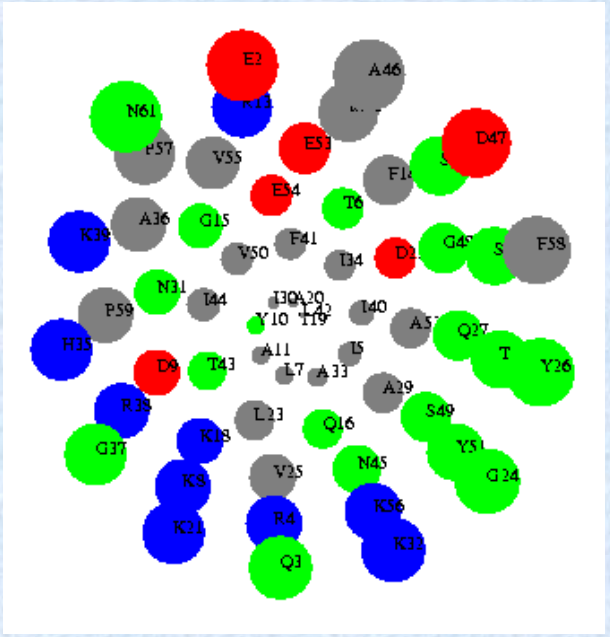
Mutants stabilize or destabilize a protein

Protein secondary structure



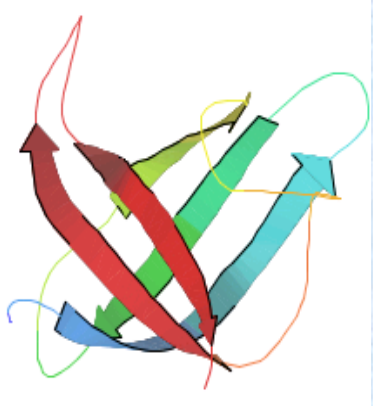
Solvent accessibility



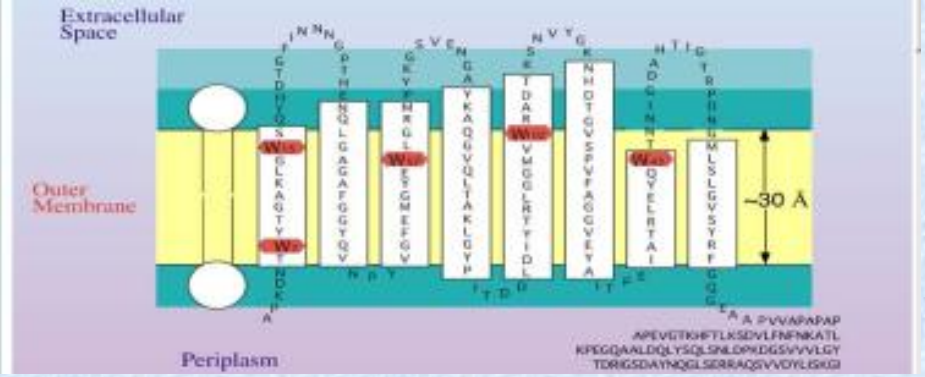


Membrane proteins

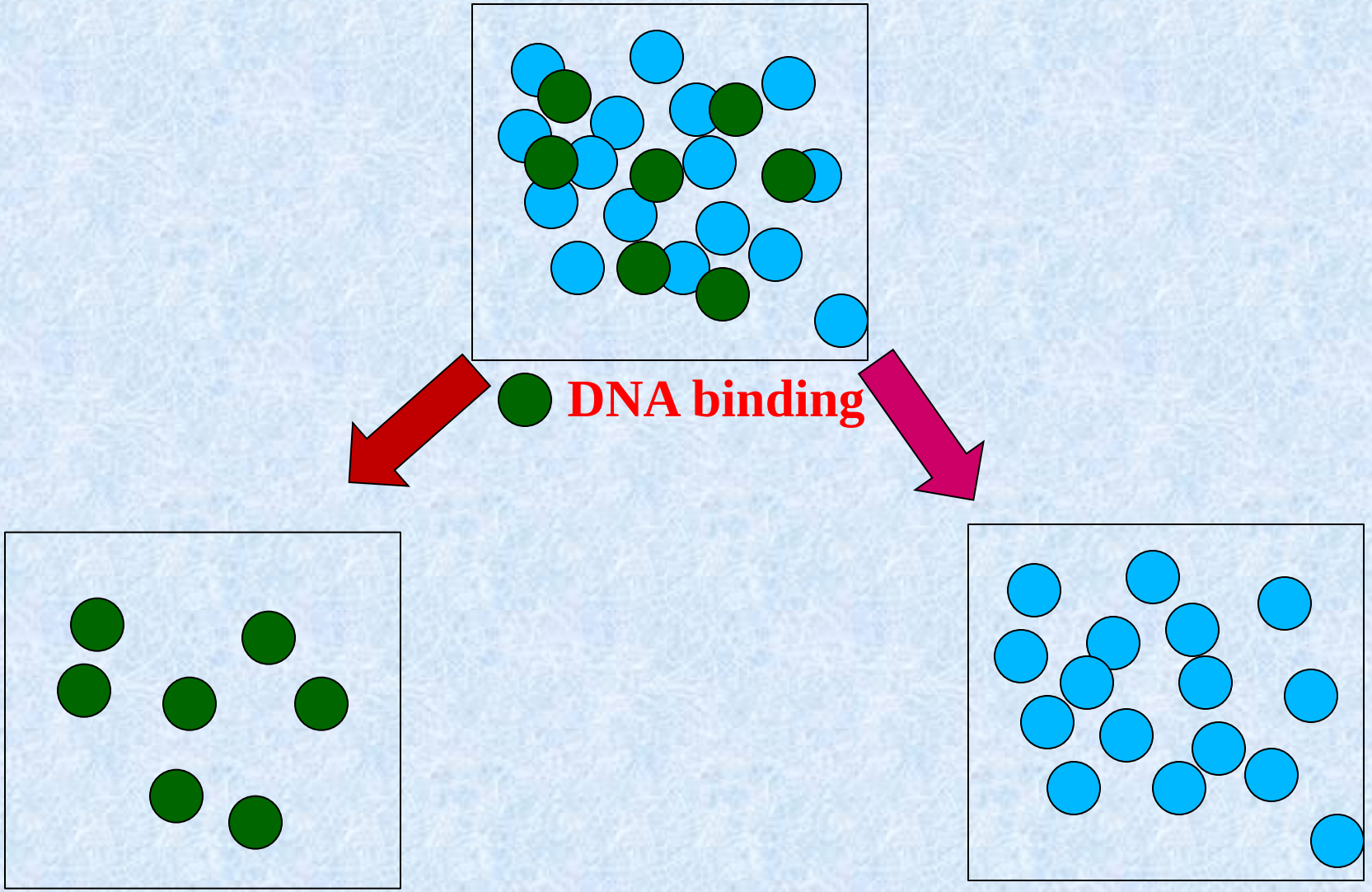
Discrimination



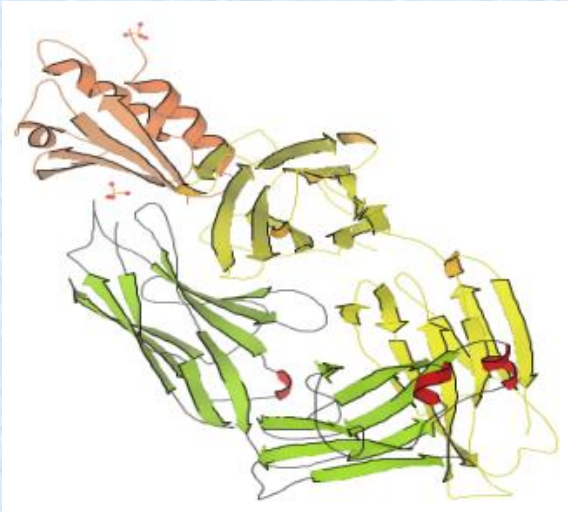
Prediction

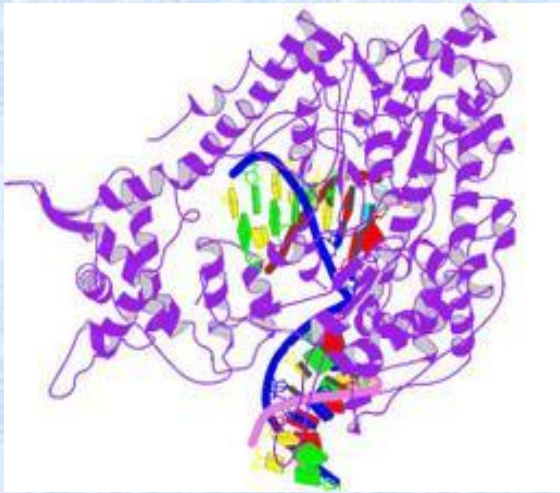


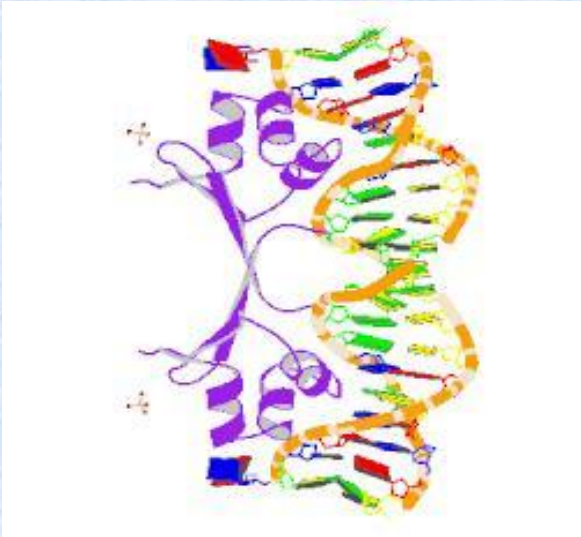
Discrimination of DNA binding proteins

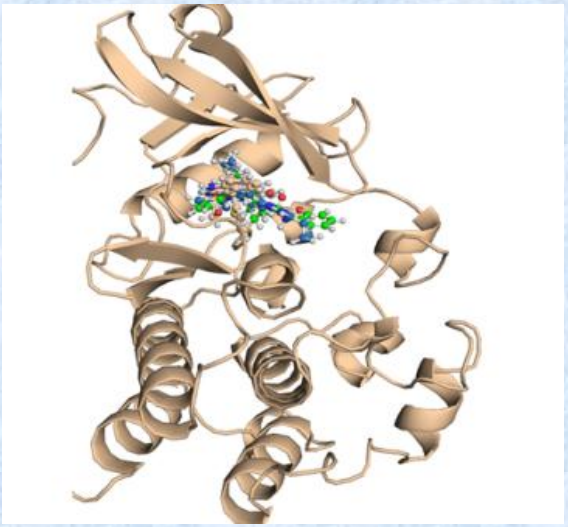


Binding sites in protein complexes









Disordered regions

Ex : 1 Free protein : 1BAM:A Complex : 3BAM:A

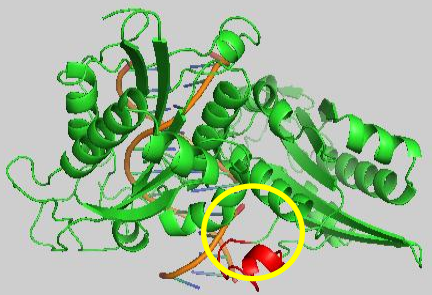
MEVEKEFITDEAKELLSKDKLIQQAYNEVKTSICSPIWPATSKTFTINNTEKNCN
GVPIKELCYTLLEDTYNWYREKPLDILKLEKKKGGPIDVYKEFIENSELKRVGME
FETGNISSAHRSMNKLLLGLKHGEIDLAIILMPIKQLAYYLTDRVTNFEELEPYF
ELTEGQPFIFIGFNAEAYNSNVPLIPKGSDGMSKRSIKKWKDKVENK

Ex : 2 Free protein : 1ES8:A Complex : 1DFM:B

MKIDITDYNHADEILNPQLWKEIEETLLKMPLHVKASDQASKVGSLIFDPVGTNQ
YIKDELVPKHWKNNIPIPKRFDFLGTDIDFGKRDTLVEVQFSNYPFLLNNTVRSE
LFHKSNMDIDEEGMKVAIIITKGHMFPASNSSLYYEQAQNQLNSLAEYNVFDVPI
RLVGLIEDFETDIDIVSTTYADKRYSRTITKRDTVKGKVIDTNTPNTRRRRKRGT
IVTY

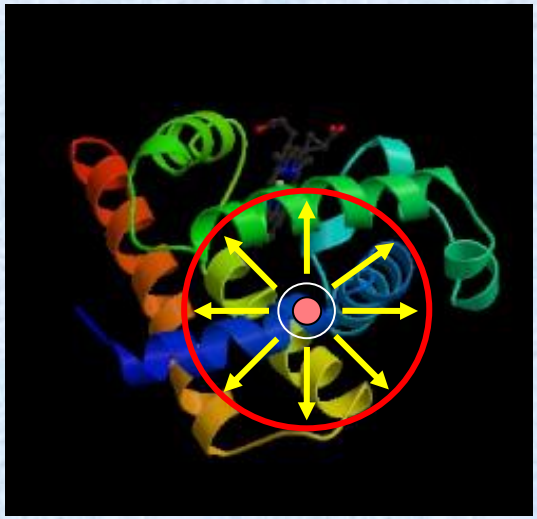
The regions which are transformed from disordered to ordered state during complex formation are indicated in red color which are encircled in the figures

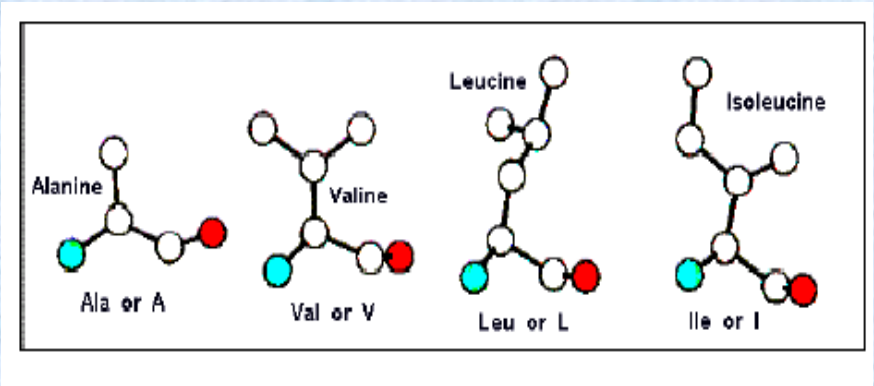
3BAM

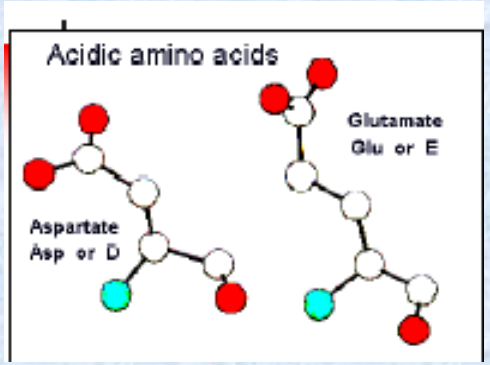


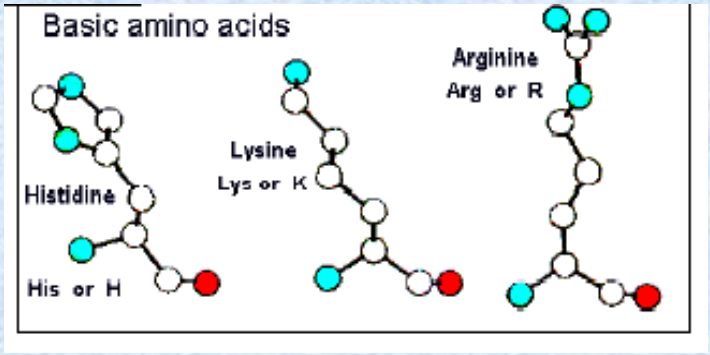
1DFM

Protein Mutant stability







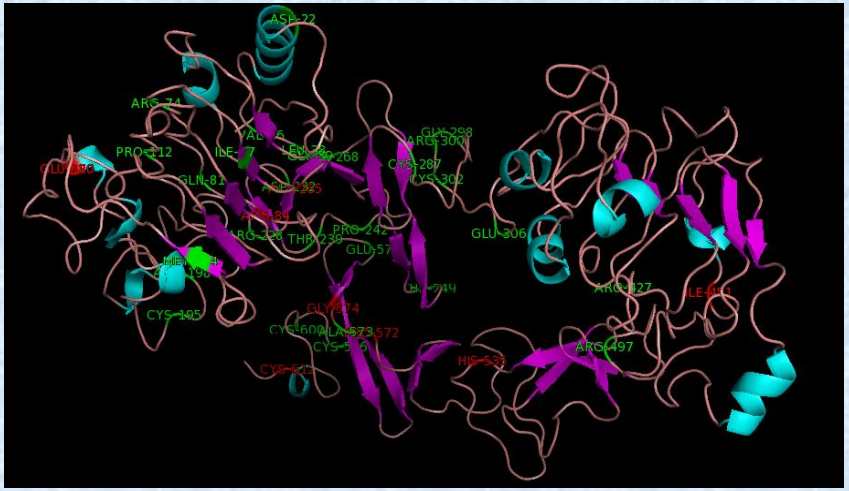


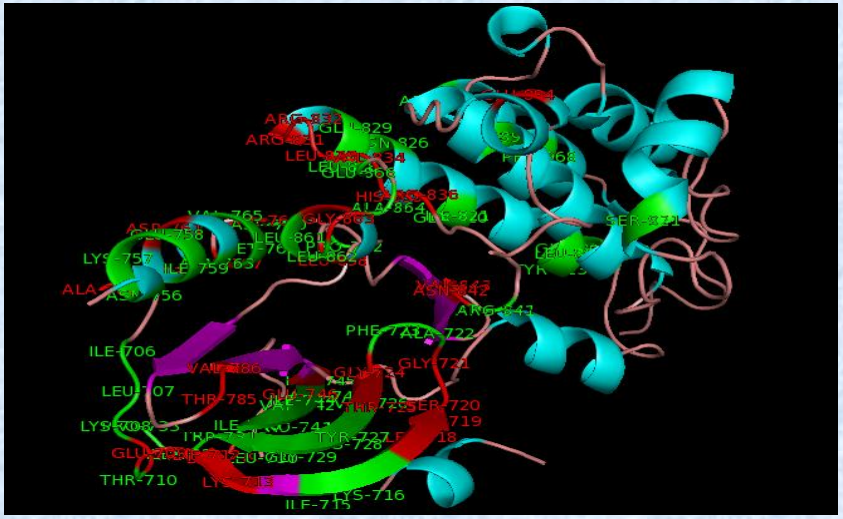
Ala → Val: collide OR well packed
Val → Ala: cavity OR freely move
Ala → Asp: electrostatic
Lys → Leu: break ion pairs

- In a protein, the **replacement** of one amino acid into others may disturb the structure, stability and function.

Hemoglobin Glu 6 → Val (sickle cell anemia); p53 mutants: cancer

Driver and passenger cancer mutations in EGFR





Research themes

Real value prediction

Stability change upon mutation

Binding affinity of protein complexes

Protein folding rates

Solvent accessibility

Construction of datasets

Non-redundant dataset

Main dataset

Test set

Blind data

Dataset is very sensitive

Feature selection

Evaluate different features

Physical/chemical Properties

Specific interactions

Amino acid composition

Residue pair preference

Motifs

Selecting important features plays a key role

Statistical methods

Regression equations

(understanding the influence of specific features)

Simple equation:

Relationship between two variables

E.g. Physiological conditions and growth of a plant

Maximum temperature

Minimum temperature

Maximum relative humidity

Minimum relative humidity

Rainfall

Rainy days per week

Statistical methods

Multiple regression technique

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots + b_n X_n$$

Y: experimental data (growth of a plant)

X_n = independent variable (physiological conditions).

In this equation, the regression coefficients (or *B* coefficients) represent the *independent* contributions of each independent variable to the prediction of the dependent variable.

Principle of least squares

Assessment

Correlation coefficient

$$r = [N\sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i)] / \sqrt{[N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2][N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2]},$$

Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \left| V_i^P - V_i^E \right|$$

Machine learning techniques

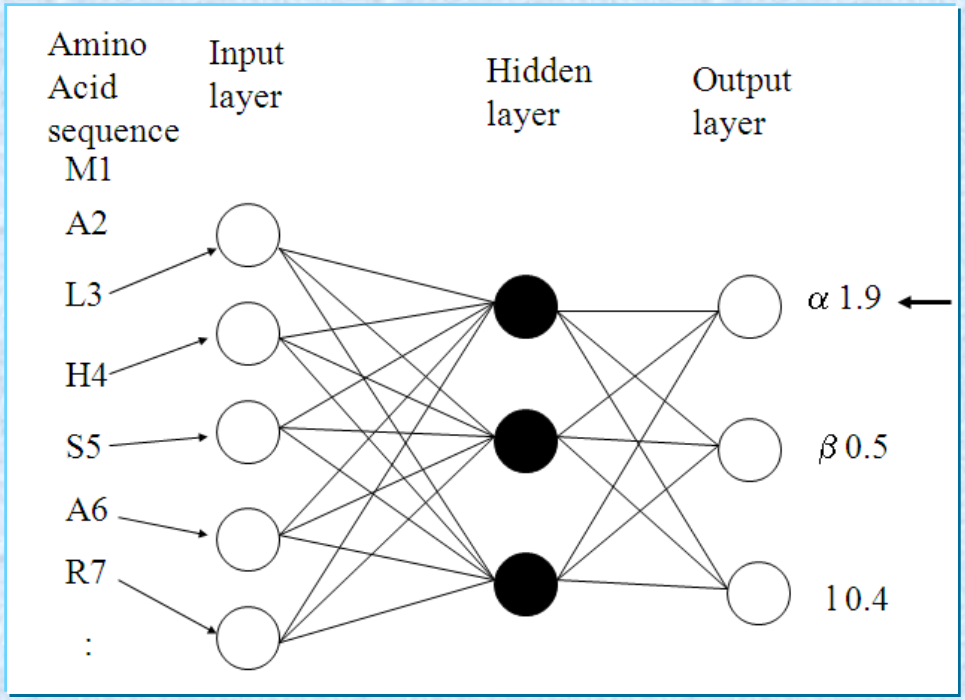
Machine learning techniques are popular in several biological applications.

This technique fits the experimental data with given input parameters and automatically selects the weights for each parameter.

Examples: Neural networks, support vector machines, classification and regression tool etc.

Neural networks

Input layer
Hidden layer
Output layer



Each node in a layer is connected to all nodes in the preceding layer.

In this case the layers are fully connected.

A weight is associated with each connecting line.

For each node an activation value is calculated based on the node's input value.

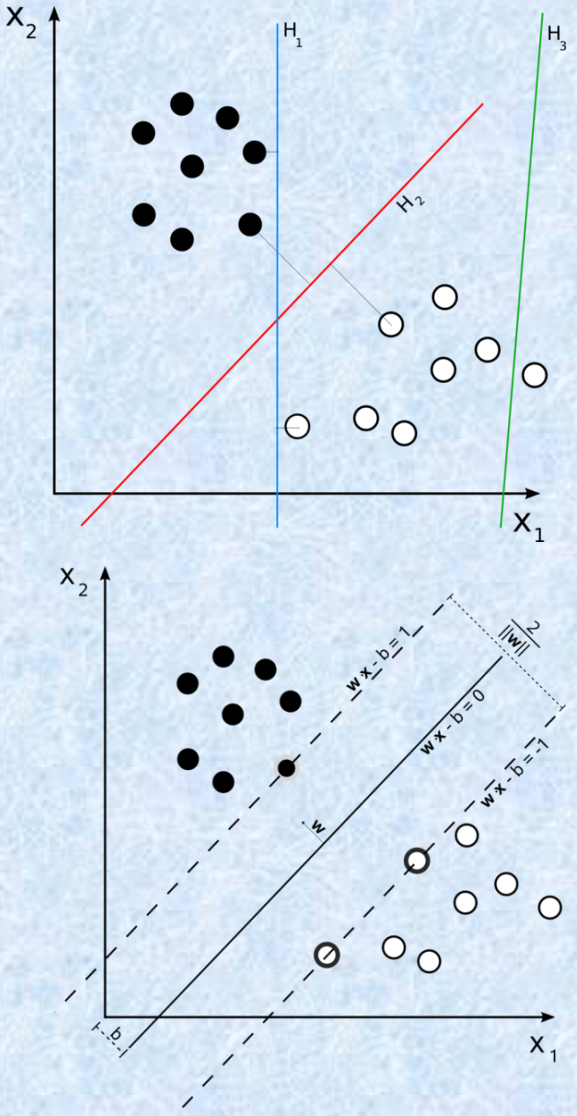
Training: Finding the values for the weights , which give high accuracy

Support vector machines

It is a learning algorithm, which from a set of positively and negatively labeled training vectors learns a classifier that can be used to classify new unlabelled test samples.

SVM learns the classifier by mapping the input training samples into a possibly high dimensional feature space, and seeking a hyperplane in this space which separates the two types of examples with the largest possible margin, i.e., distance to the nearest points.

If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin.



Assessment measures

2-groups classification

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Experimental	Predicted +	Predicted -
+	TP	FN
-	FP	TN

TP: True positive
FP: False positive
TN: True negative
FN: False negative

```
Experimental:   NNNNNBBBBNNBNNNNBBBNNNNNBNNNBNNNBBN
Predicted:      NNNNNBBBBNNNNBBBBNNNNNNNNNNNNBBBNNN
```

Assessment with different measures is necessary

Discrimination results: β-barrel membrane proteins

Datasets: 208 TMβs and 879 non-outer membrane

(673 globular and 206 α-helical membrane; nTMβ) proteins

Method	Parameters	TMβ	nTMβ	Accuracy (%)	
Statistical	AA	89	79	83	**Bioinformatics, 2005**
Statistical	Pairs	95	79	86	**CBAC, 2005**
Statistical	Motifs	96	86	90	**BPC, 2005**
SVM	18 AA + 10 pairs	91	95	94	**Bioinformatics, 2005**
NN	49 properties	81	98	94	**BBA, 2006**
RBF	PSSM profiles	89	98	96	**CBAC, 2008**

18 AA: Except Ala and Glu; 10 pairs: QA, DF, DA, KK, EF, NK, DR, YN, FF and LI

Statistical: High sensitivity (correctly identifying TMBs)

Machine learning: High specificity (correctly excluding non-TMBs)

Machine learning techniques: Cautions

Over-fitting

Number of data

Number of features

Cautious about over-fitting of parameters

Validation procedures

1. Self assessment (back-check)
2. Training and test set
3. n-fold cross validation
4. Jack-knife test
5. Split sampling

Systematic validation procedure shows the performance

Comparison with other methods

Different datasets

Measures

Validations

New data

Systematic comparison shows the superiority

Constructing web servers

Beneficial to other researchers

- User friendly
- Effective algorithms
- Reliable results

Gene and Genome

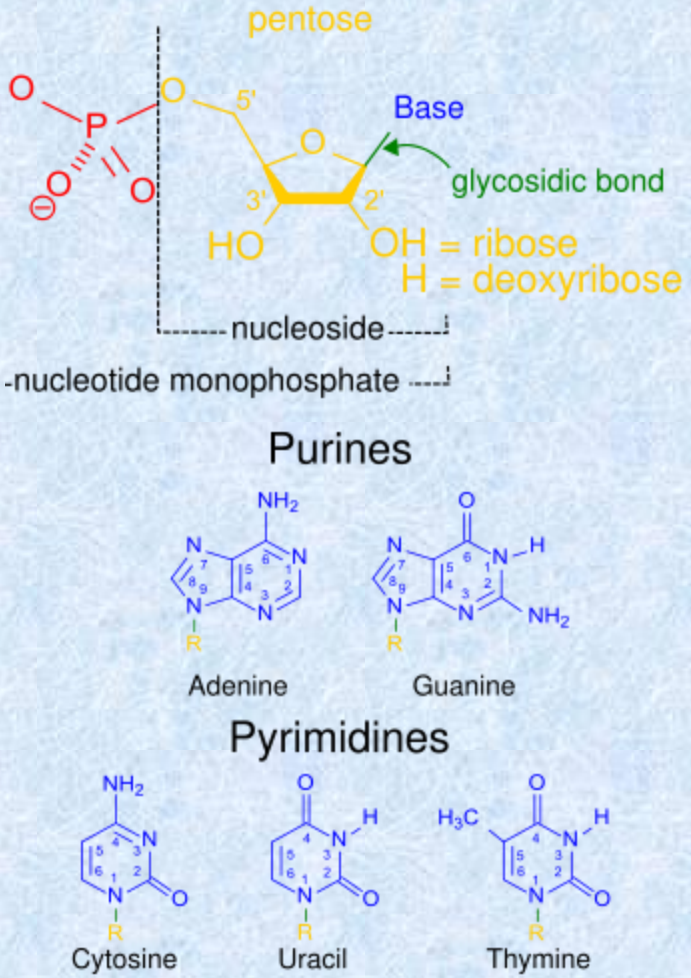
A gene is a unit of heredity in a living organism. It is normally a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism.

In most living organisms genetic information is stored in the molecule deoxyribonucleic acid, or DNA.

DNA is made and resides in the nucleus of living cells.

DNA gets its name from the sugar molecule contained in its backbone (deoxyribose);

Four different nucleotide bases occur in DNA: **adenine (A), cytosine (C), guanine (G), and thymine (T).**



Genome is the entirety of an organism's hereditary information.

Human genome

Human genome occupies a total of just over 3 billion DNA base pairs.

Human genome contains ca. 23,000 protein-coding genes.

A personal genome sequence is a complete sequencing of the chemical base pairs that make up the DNA of a single person.

Because medical treatments have different effects on different people because of genetic variations the analysis of personal genomes may lead to personalized medical treatment.

The genome of Kim (Korea) had 1.58 million alterations

Six out of 10,000 DNA bases are unique to Koreans


Homo sapiens chromosome 1 genomic contig, GRCh37

NCBI Reference Sequence: NT_077402.2

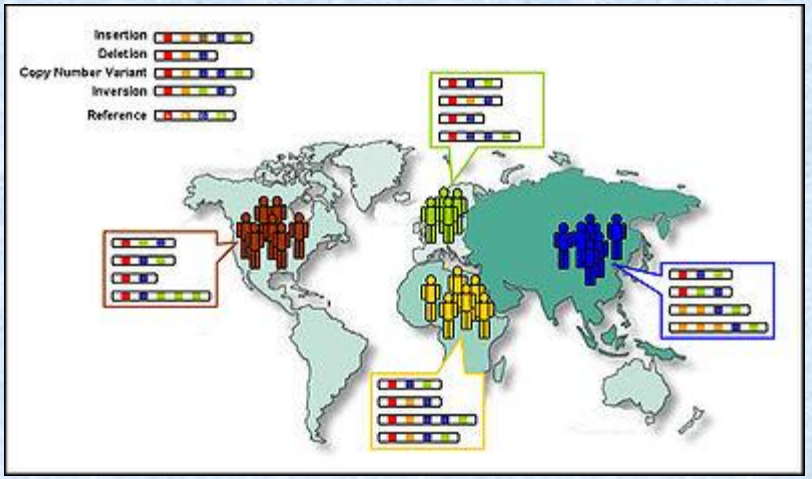
>gi|224514618|ref|NT_077402.2| Homo sapiens chromosome 1 genomic contig, reference primary assembly

TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCCCTAACCCCTAACCCTAACCCTAACC
CTACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCG
CCCGCCCGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGG
CGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCG
GCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAG
AGAGGCGCGCCGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGC
AGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCG
TGGCGCAGGCGCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGC
ACGGCGCCGGGCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCAC
TACAGGACCCGCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCT
CTCTTGCTTAGAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGT
ATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCA
GCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACAGTACTGGCGGATTATAG
GGAAACACCCGGAGCATATGCTGTTTGGTCTCAGTAGACTCCTAAATATGGGATTCCTGGGTTTAAAAGT
AAAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAGCAA
TGTCTAGGAATGCCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCC
TGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGATGATT
TTGCTGCATGGCCGGTGTTGAGAATGACTGCGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAG
TTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTT
CTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCCAGACTTCCCGTGTCCTTTCCACC
GGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTG
CTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAG
TGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATACCACAACCAGGCATAG
GGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCTAGGCCAGTAAGTAGTGCTT
GTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTG
CTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGT
CTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGC
CGCAGGCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCC
CCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGA
GCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCAGAGGTGAGAGGA
GAGTAGACAGTGAGTGGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCTGGAGCCCTGCTGCTTGCGGT
GGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCA
AACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCCAAGAAAGGCTC
TGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCG
ACAGACAGGGGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCTTAGCCTCTGTTCCCACGAAG
GCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCC
TGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATT
GCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTC
TCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT

1000 genomes project

The 1000 Genomes Project, launched in January 2008, is an international research effort to establish the most detailed catalogue of human genetic variation.

Sequence the genomes of at least **one thousand** anonymous participants from a number of different ethnic groups within the next three years, using newly developed technologies which are faster and less expensive.



An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aim build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1, individuals from 14 populations, constructed using a combination of low-coverage whole-genome and ex sequencing. By developing methods to integrate information across several algorithms and diverse data sources, provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry diffe profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiat which is further increased by the action of purifying selection. We show that evolutionary conservation and co consequence are key determinants of the strength of purifying selection, that rare-variant load varies substanti across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved s such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common low-frequency variants in individuals from diverse, including admixed, populations.

Nature, 2010

Cancer genome project

The Cancer Genome Project aims to identify sequence variants/mutations critical in the development of human cancers.

Cancer Genome Project represents an effort to improve cancer diagnosis, treatment, and prevention through a better understanding of the molecular basis of this disease.

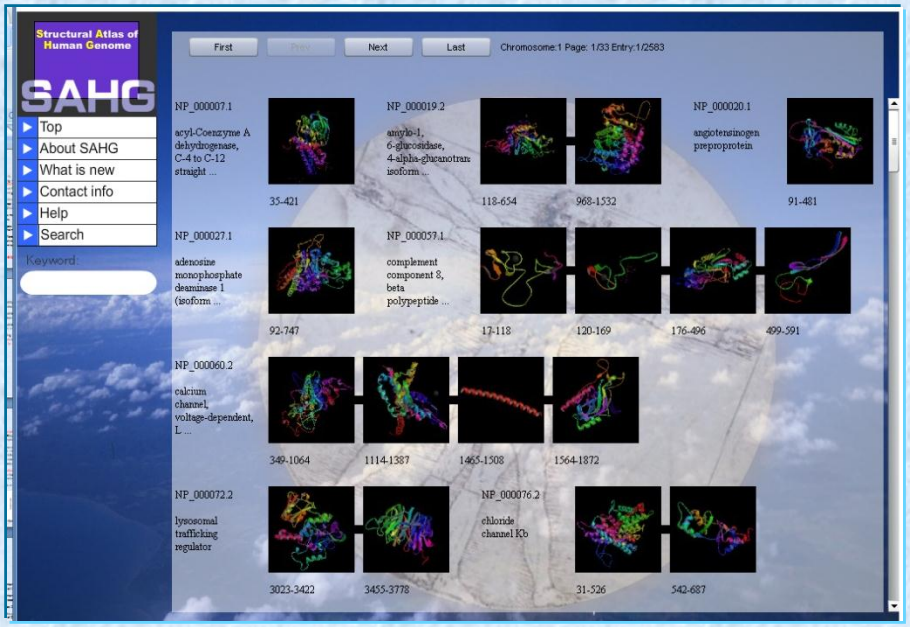
The Cancer Genome Project combines knowledge of the human genome sequence with high throughput mutation detection techniques.

Human genome: Bioinformatics

Structural Atlas of Human Genome (Japan)

Structural and Functional Annotation

- **Protein-protein complexes**
- **Protein-DNA complexes**
- **Protein-RNA complexes**
- **Protein-small molecule complexes**
- **Membrane proteins**
- **Subcellular localization**
- **Transporters**
- **Receptors**
- **Enzymes**
- **Coiled coil proteins**



Open problems

Annotate genomes based on structure and function of proteins

Cross-genome analysis

Personalized medicine