SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

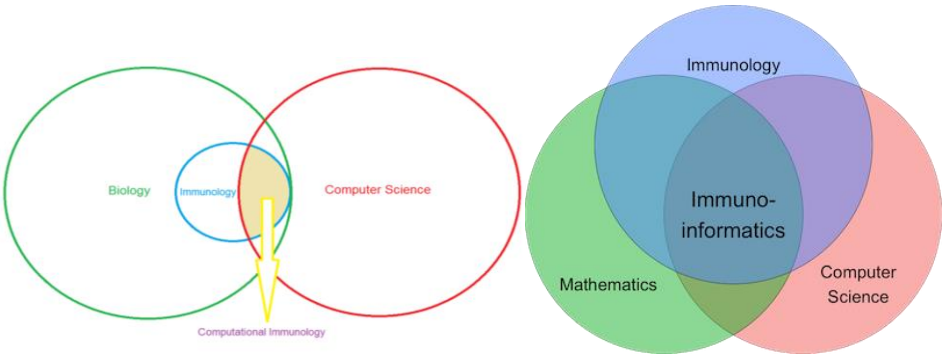
DEPARTMENT OF BIOINFORMATICS

UNIT – I -IMMUNOINFORMATICS&COMPUTATIONAL VACCINOLOGY-SBIA5202

UNIT I INTRODUCTION TO IMMUNOINFORMATICS

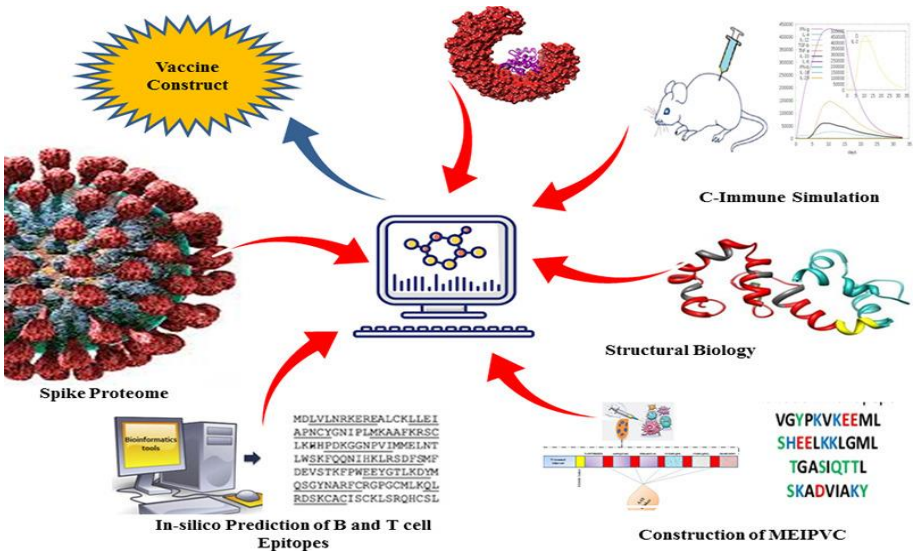
Definition of Immunoinformatics

- Immunoinformatics involves the application of computational methods to immunological problems.

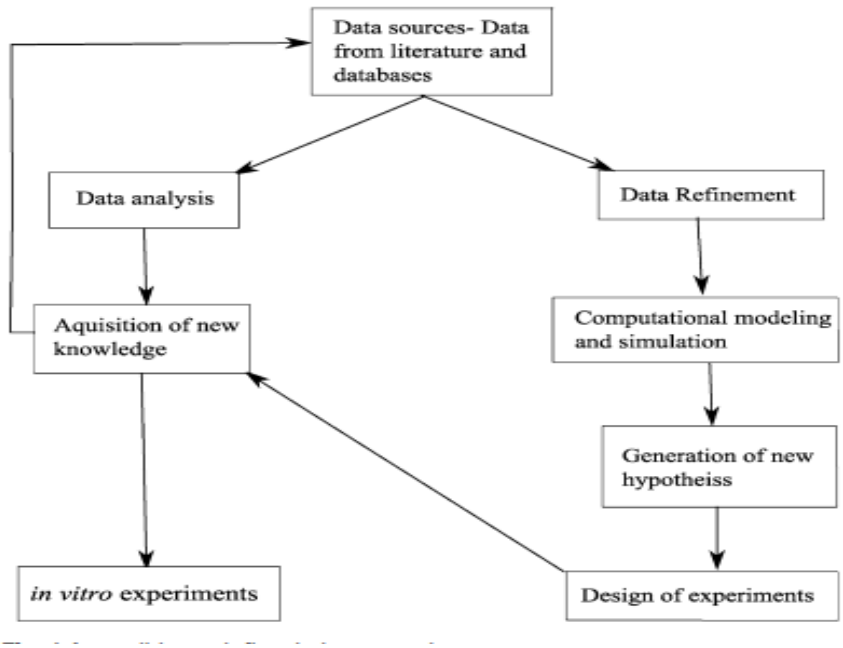


- Prediction of B- and T-cell epitopes has long been the focus of immunoinformatics, given the potential translational implications, and many tools have been developed.
- With the advent of next-generation sequencing (NGS) methods, an unprecedented wealth of information has become available that requires more-advanced immunoinformatics tools.

A large volume of data relevant to immunology research has accumulated due to sequencing of genomes of the human and other model organisms. At the same time, huge amounts of clinical and epidemiologic data are being deposited in various scientifi c literature and clinical records. This accumulation of the information is like a goldmine for researchers looking for mechanisms of immune function and disease pathogenesis. Thus the need to handle this rapidly growing immunological resource has given rise to the fi eld known as immunoinformatics. Immunoinformatics, otherwise known as computational immunology, is the interface between computer science and experimental immunology. It represents the use of computational methods and resources for the understanding of immunological information. It not only helps in dealing with huge amount of data but also plays a great role in defi ning new hypotheses related to immune responses. This chapter reviews classical immunology, different databases, and prediction tool. Further, it briefl y describes applications of immunoinformatics in reverse vaccinology, immune system modeling, and cancer diagnosis and therapy. It also explores the idea of integrating immunoinformatics with systems biology for the development of personalized medicine. All these efforts save time and cost to a great extent.



All the genes and proteins taking part in immune responses are referred to as “immunome,” and it excludes genes and proteins that are expressed in cell types other than in immune cells. All immune reactions due to interaction between host and antigenic peptides are referred to as “immunome reactions,” and their study is called as “immunomics”. Like genomics and proteomics, immunomics is a new discipline, which uses high-throughput techniques to understand immune system mechanism]. Figure shows work flow in immunomics. This chapter describes various available information regarding classical immunology, different immunomic databases, B and T cell epitope prediction tools and software, and applications of immunoinformatics.



Overview of the Immune system

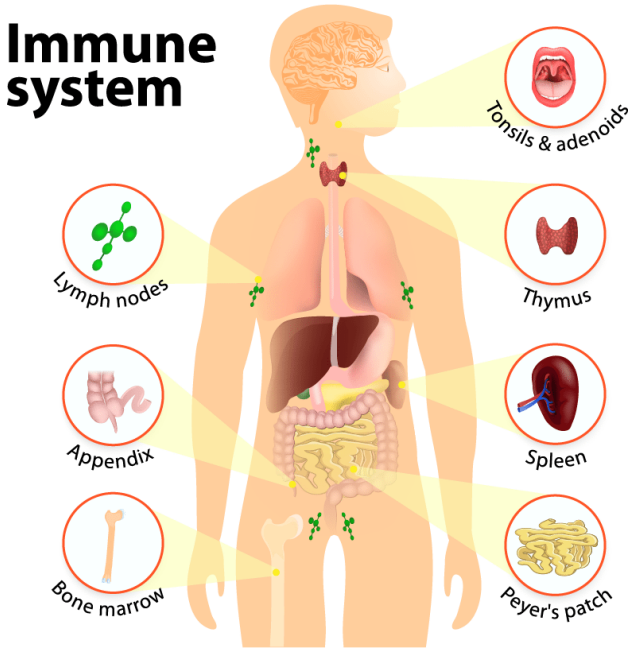
The immune system is your body's defense mechanism against disease and infection, and is responsible for targeting and destroying substances that it recognizes as foreign or different from normal, healthy tissues in the body.

The Components

The primary components of the immune system are:

- **The tonsils and the thymus:** These are responsible for producing antibodies, which are some of the combatants against foreign invaders in the body.
- **Lymphatic system:** Made up of lymph nodes and vessels, this is a network that carries lymph fluid, nutrients and waste material between the body's tissues and the bloodstream. The lymph nodes filter lymph fluid as it flows through them, trapping bacteria, viruses and other invaders. These invaders are then destroyed by special white blood cells called lymphocytes.
- **Bone marrow:** This is the soft tissue found primarily inside the long bones of the arms, legs, vertebrae and the pelvic bones in the body. It's made of red marrow, which produces red and white blood cells along with platelets and yellow marrow. Yellow marrow contains fat and connective tissue and helps produce some white blood cells.

- **Spleen:** The spleen filters the blood by removing old or damaged cells or platelets. It also helps the immune system by destroying bacteria and other invaders.
- **White blood cells:** Made in the bone marrow, these cells protect your body from infection. If an infection develops, white blood cells attack and destroy the organism causing it, whether it's bacteria, a virus or another organism.

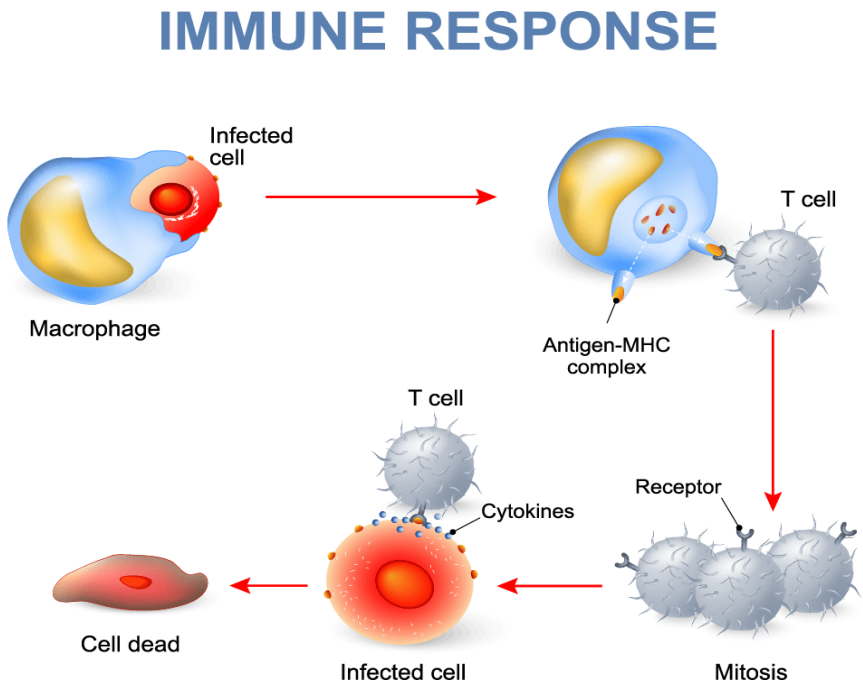


How The Immune System Works

Key cells in your immune system, lymphocytes known as B and T cells, help destroy invaders within the lymphatic system. Their process goes as follows:

- After T cells develop in the thymus, all immune system cells gather in the lymph nodes and spleen.
- First, antigens are ingested and partially digested. They are then presented to helper T cells by other cells called macrophages. This activates the T cells to release hormones that help B cells develop.
- These hormones, plus the recognition of further antigens, change the B cell into a plasma cell that produces antibodies—these antibodies may come in several types and will fit the antigen like a lock fits a key, thus rendering the antigen itself harmless.

- Helper T cells also aid in development of cytotoxic T cells, which can directly kill antigens. In addition, memory T cells are produced so that any re-exposure to the same antigen will produce a quicker and more effective response.



Immune Cells

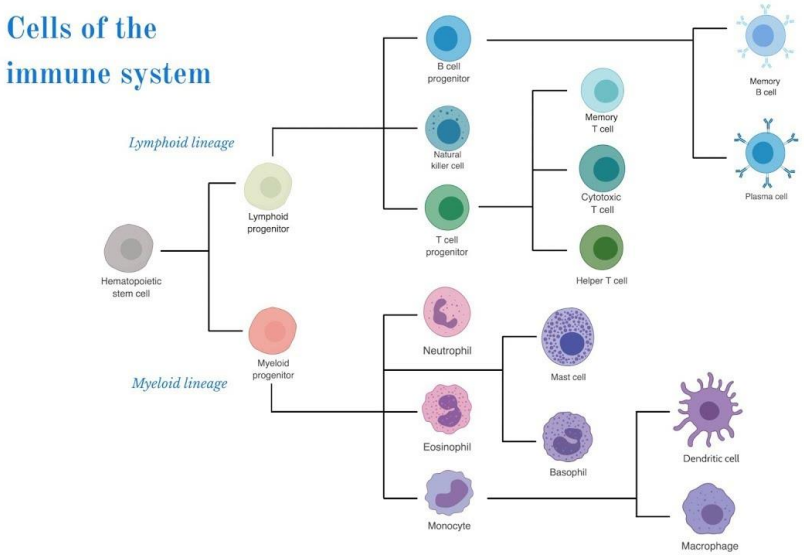
Many cells work together as part of the innate (non-specific) and adaptive (specific) immune system. See the module "Innate vs. Adaptive Immune Response" for more information on innate and adaptive immune response. Immune cells are sometimes called white blood cells or leukocytes.

Granulocytes are a type of leukocyte that contain granules in their cytoplasm containing enzymes. Neutrophils, basophils and eosinophils are types of granulocytes. Neutrophils are considered the first responders of the innate immune system. Neutrophils and macrophages circulate though the blood and reside in tissues watching for potential problems. Both cells can "eat" bacteria, as well as communicate with other immune cells if an issue arises.

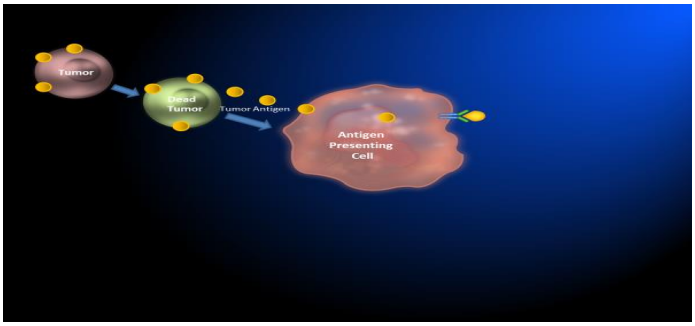
Cells of the adaptive immune system (also called immune effector cells) carry out an immune function in response to a stimulus. Natural killer T lymphocytes and B lymphocytes are examples of effector cells. For example, activated T lymphocytes destroy pathogens via cell-mediated

response. Activated B cells secrete antibodies that aid in mounting an immune response. Effector cells are involved in the destruction of cancer.

Non-effector cells are antigen-presenting cells (APCs), such as dendritic cells, regulatory T cells, tumor-associated macrophages and myeloid-derived suppressor cells. Non-effector cells cannot cause tumor death on their own. Non-effector cells prevent the immune action of the effector cells. In cancer, non-effector cells allow tumors to grow.



Component	**General Description**
Antigen	Any substance that is able to cause an immune response in the body. Examples include bacteria, chemicals, toxins, viruses and pollen. Cells in the body, as well as cancer cells, have antigens that can cause an immune response. Tumor cells originate from normal cells, but they make non-self antigens and “neoantigens” which are derived from mutated self

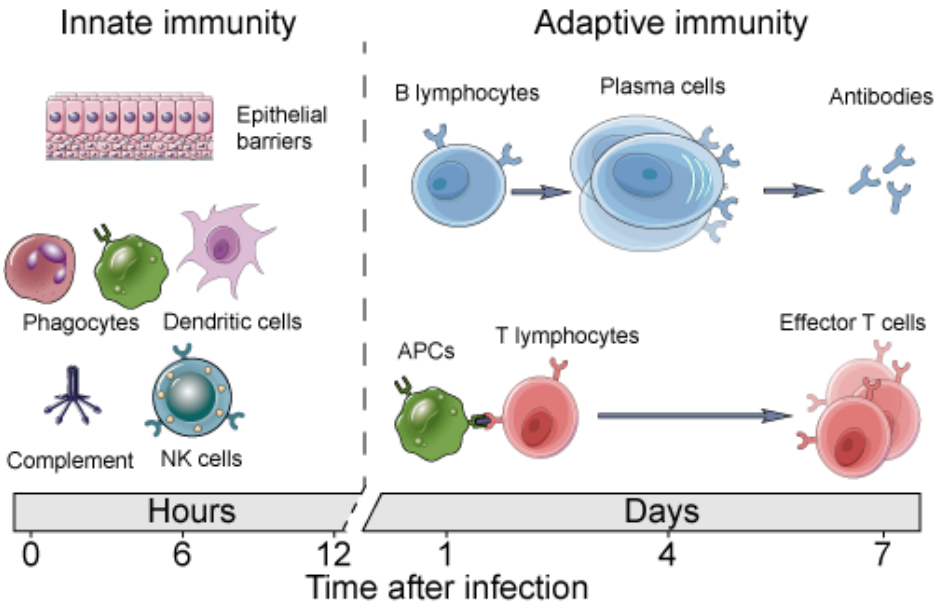
	protein. Tumor antigens can trigger adaptive immunity.
Antigen presenting cell (APC)	Cells, such as macrophages, dendritic cells and B cells, that can process protein antigens into peptides. These peptides can then be presented (along with major histocompatibility complex) to T-cell receptors on the surface of the cell.  **Figure.** Tumor Antigens Presented on Antigen Presenting Cell Source: Asim Amin, MD, Levine Cancer Institute, Atrium Health
Antibody (Ab)	Special proteins created by white blood cells that can kill or weaken infection-causing organisms. Antibodies travel through the blood stream looking for specific pathogens. The body can create new antibodies in response to new pathogens or vaccines. Also referred to as immunoglobulin (Ig).
Basophil	A basophil is a type of phagocytic immune cell that has granules. Inflammation causes basophils to release histamine during allergic reactions.

B lymphocyte	A B lymphocyte is a type of white blood cell that develops in the bone marrow and makes antibodies.
Memory B cell	B cells that are long lived and remember past antigen exposure.
Plasma B cell	Activated B cells that produce antibodies. Only one type of antibody is produced per plasma B cell.
Cytokine	A type of protein that impacts the immune system by either ramping it up or slowing it down. Cytokines can occur naturally in the body or be produced in a laboratory. Interferon-alpha2b is a cytokine produced in a laboratory (using recombinant DNA technology) and is used in the treatment of malignant melanoma.
Dendritic cell	Dendritic cells are antigen-presenting cells (APCs). Antigen is combined with major histocompatibility complex and presented on a dendritic cell to active T and B lymphocytes.
Eosinophil	An eosinophil is a type of immune cell (leukocyte, or white blood cell). They help fight infection or cause inflammation.
Granulocyte	Granulocytes (including eosinophils, neutrophils and basophils) are a type of white blood cell that releases toxic materials, such as antimicrobial agents, enzymes, nitrogen oxides and other proteins, during an attack from a pathogen.
Human leukocyte	Human version of the major histocompatibility complex (MHC).

antigens	The MHC complex is a family of 200+ genes categorized into three classes: I, II, III. Class I genes make proteins that are located on the surface of almost all cells. Class II genes are located on the surface of immune cells. Class III genes are also involved with the immune system and inflammation.
Natural killer (NK) cell	The primary effector cell of innate immunity; the first responders of the immune system. They interact with signals from other cells (activating and inhibitory).
T lymphocyte (also called T cell)	Type of white blood cell that is involved with the immune system. T lymphocytes mature in the thymus and differentiate into cytotoxic, memory, helper and regulatory T cells. CAR T-cell therapy uses T cells obtained from a patient's own blood to fight cancer. The T cells are grown and modified in a lab to include special receptors (chimeric antigen receptor) that can recognize and attack cancer cells.
Cytotoxic T cell	Cytotoxic T cells are the primary effector cells of adaptive immunity. Activated cytotoxic T cells can migrate through blood vessel walls and non-lymphoid tissues. They can also travel across the blood brain barrier. Cytotoxic T cells are activated by cytokines. They can attach to cancer cells and kill them.

Memory T cell	Derived from activated cytotoxic T cells, memory T cells are long-lived and antigen-experienced. One memory T cell can produce multiple cytotoxic T cells. After activated cytotoxic T cells attack the pathogen, the memory T cells hang around to mitigate any recurrence.
Helper T cell	Helper T cells secrete cytokines that help B cells differentiate into plasma cells. These cells also help to activate cytotoxic T cells and macrophages.
Regulatory T cell	Regulatory T cells (or Tregs) help to suppress the immune system.
Lymphocyte	Lymphocytes are immune cells found in the blood and lymph tissue. T and B lymphocytes are the two main types.
Macrophage	Macrophages are large white blood cells that reside in tissues that specialize in engulfing and digesting cellular debris, pathogens and other foreign substances in the body.
Major histocompatibility complex (MHC)	MHC is a group of genes that code for proteins on the cells of the immune system. Referred to as the human leukocyte antigen (HLA) system in humans.
Mast cell	Mast cells release histamine and help to get rid of allergens.
Monocyte	Large white blood cells that reside in the blood stream that specialize in engulfing and digesting cellular debris, pathogens and other foreign substances in the body. Monocytes become macrophages.
Myeloid-derived	When immature myeloid cells cannot differentiate into mature myeloid

suppressor cells	cells, due to conditions like cancer, expansion of myeloid-derived suppressor cells occurs, and the T-cell response can be suppressed.
Neutrophil	A type of white blood cell, granulocyte, and phagocyte that aids in fighting infection. Neutrophils kill pathogens by ingesting them.
Phagocytes	Phagocytes eat up pathogens by attaching to and wrapping around the pathogen to engulf it. Once the pathogen is trapped inside the phagocyte, it is in a compartment called a phagosome. The phagosome will then merge with a lysosome or granule to form a phagolysosome, where the pathogen is killed by toxic materials, such as antimicrobial agents, enzymes, nitrogen oxides or other proteins.



The Innate vs. Adaptive Immune Response

The first line of defense against non-self pathogens is the innate, or non-specific, immune response. The innate immune response consists of physical, chemical and cellular defenses against pathogens.

The main purpose of the innate immune response is to immediately prevent the spread and movement of foreign pathogens throughout the body.

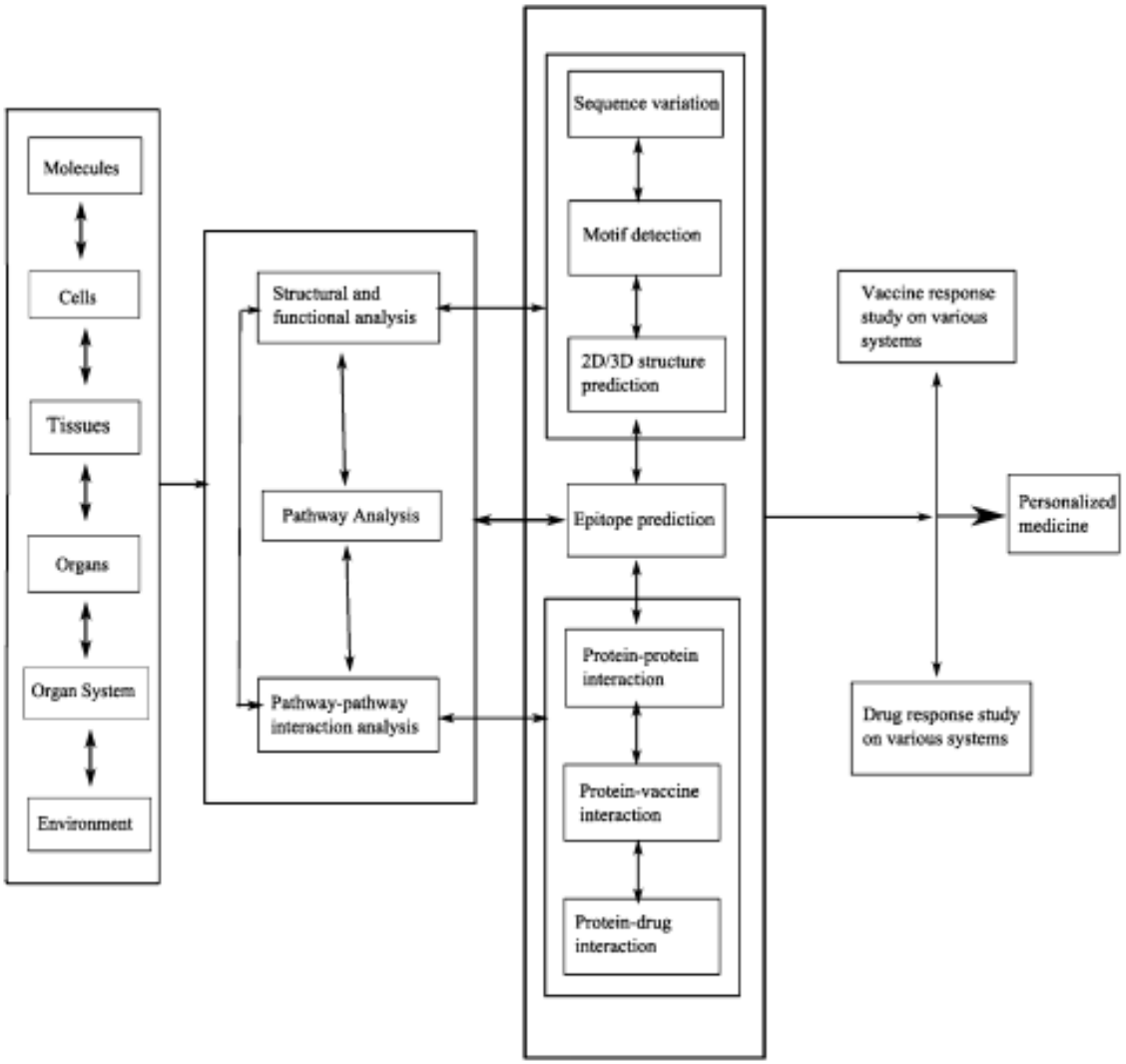
The second line of defense against non-self pathogens is called adaptive immune response. Adaptive immunity is also referred to as acquired immunity or specific immunity and is only found in vertebrates. The adaptive immune response is specific to the pathogen presented. The adaptive immune response is meant to attack non-self pathogens but can sometimes make errors and attack itself. When this happens, autoimmune diseases can develop (e.g., lupus, rheumatoid arthritis).

The hallmark of the adaptive immune system is clonal expansion of lymphocytes. Clonal expansion is the rapid increase of T and B lymphocytes from one or a few cells to millions. Each clone that originates from the original T or B lymphocyte has the same antigen receptor as the original and fights the same pathogen.

While the innate immune response is immediate, the adaptive immune response is not. However, the effect of the adaptive immune response is long-lasting, highly specific, and is sustained long-term by memory T cells.

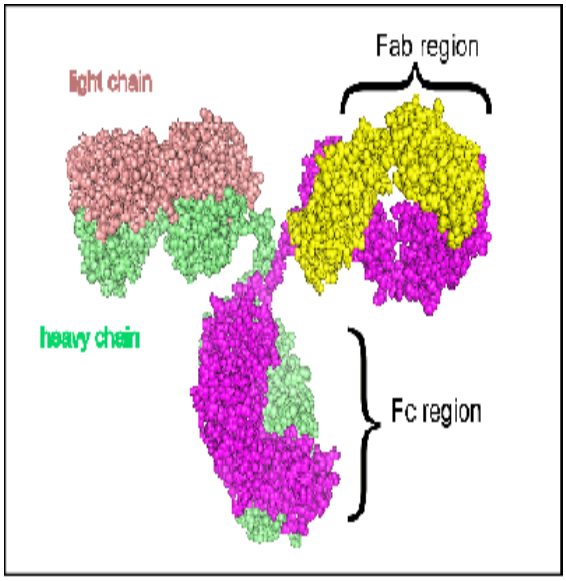
	Line of Defense	**Timeline**	**Cells**	**Antigen Dependency**	**Examples**
Innate (non-specific)	First	Immediate response (0 - 96 hours)	Natural killer cells, macrophages, neutrophils, dendritic cells, mast cells, basophils, eosinophils	Independent	Skin, hair, cough, mucous membranes, phagocytes, granulocytes
Adaptive (specific)	Second	Long term (>96 hours)	T and B lymphocytes	Dependent	Pus, swelling, redness, pain, T and B

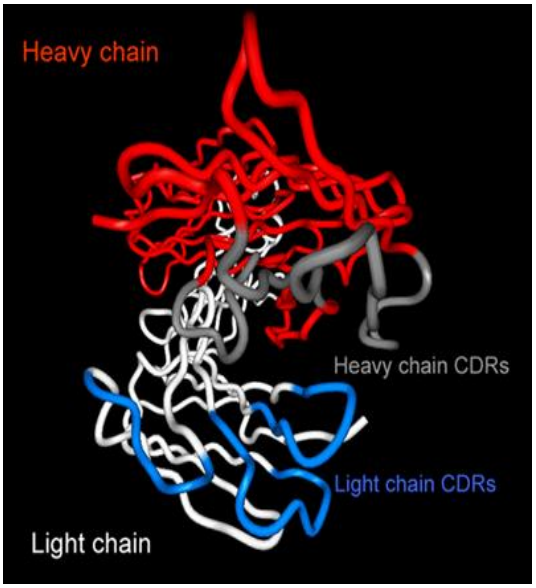
					lymphocyte response



- Antibodies, or immunoglobulines (Ig), are Y shaped proteins complexes composed of four polypeptide chains, two identical light chains and two identical heavy chains.
- Each light chain is bound to a heavy chain via disulfide bridges to form a heterodimer.
- Two identical heterodimers are linked to each other by disulfide bridges to form the basic antibody molecule (Fig).

- The constant domains of the heavy chain (Fc) determine the biological function of the antibody creating five major classes of antibodies found in higher verterbrates (IgG, IgA, IgD, IgE and IgM).
- The variable domains (Fab), however, are involved in antigen binding. Within the Fab regions there are six hypervariable loops (3 on the surface of the light chain and 3 on the surface of the heavy chains) that directly interact with the antigen named the Complementarity Determining Regions (CDRs) (Fig).





Brief introduction to immunoformatics and computational vaccinology

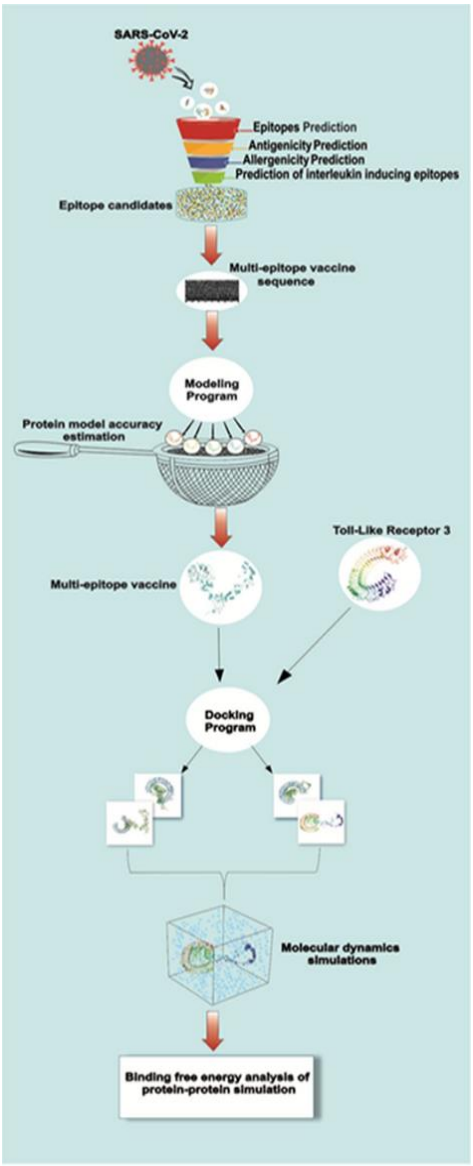
Immunology studies produce data in colossal quantities. Also, with proteomics and genomics projects, extensive screening of pathogens and/or pathogen–host interaction, it has become increasingly necessary to store, manage and analyze these data, hence the birth of immunoinformatics. Immunoinformatics deals with computational techniques and resources used to study the immune functions. Statistical, computational, mathematical and biological knowledge and tools are applied in immunoinformatics in order to accurately and specifically store, and analyze data concerning the immune system and its functions.

To handle evidence diversity, immunoinformatics uses tools that cut across several aspects of bioinformatics such as creation and management of databases, use and definition of both structural and functional signatures and the formation and application of predictive tools. These strategies can

synergize toward a better understanding of the immune system of both man and animals and fight against some less predictable pathogenesis. The complex nature of vertebrates' immune system, the variable nature of pathogens and environmental antigens coupled with the multi-regulatory pathways show that colossal quantities of data will be needed to unveil how the human immune systems work. Conventionally, much cannot be achieved based on the complexity of the immune system and the virulent antigen but with the application of computational vaccinology, researches on vaccine design have been made easier, accurate and specific. Applying immunoinformatics in disease study requires the knowledge of disease pathogenesis, the immune system dynamics, and computational vaccinology, painstaking searches of the database, sequence comparison, structural modeling as well as motif analysis. These methods can go a long way in analyzing the pathogenesis of a disease and identification of vaccine candidates.

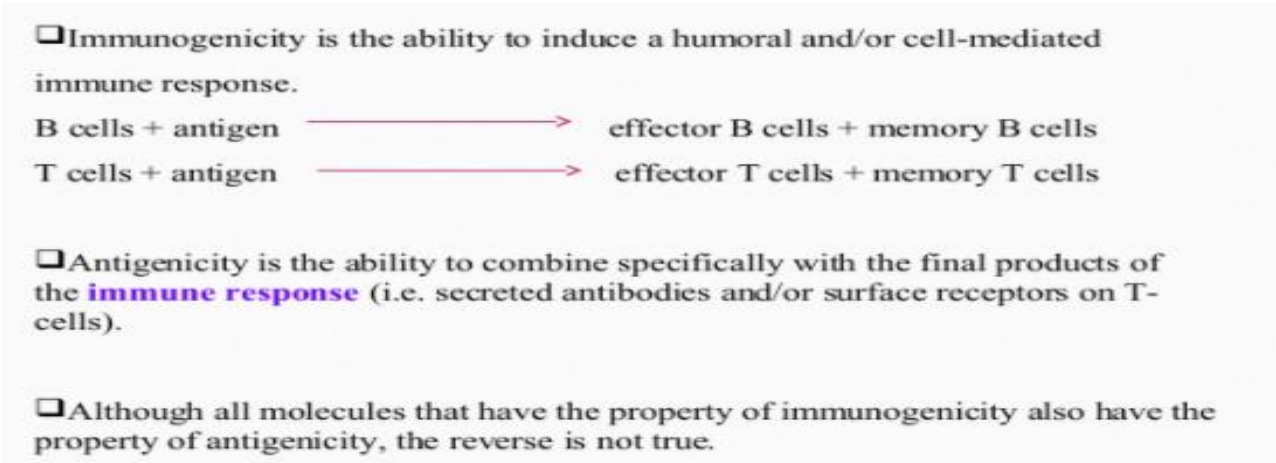
Immunoinformatic design of a COVID-19 subunit vaccine using entire structural immunogenic epitopes of SARS-CoV-2

Coronavirus disease 2019 (COVID-19) is an acute pneumonic disease, with no prophylactic or specific therapeutical solution. Effective and rapid countermeasure against the spread of the disease's associated virus, SARS-CoV-2, requires to incorporate the computational approach. In this study, we employed various immunoinformatics tools to design a multi-epitope vaccine polypeptide with the highest potential for activating the human immune system against SARS-CoV-2. The initial epitope set was extracted from the whole set of viral structural proteins. Potential non-toxic and non-allergenic T-cell and B-cell binding and cytokine inducing epitopes were then identified through a priori prediction. Selected epitopes were bound to each other with appropriate linkers, followed by appending a suitable adjuvant to increase the immunogenicity of the vaccine polypeptide. Molecular modelling of the 3D structure of the vaccine construct, docking, molecular dynamics simulations and free energy calculations confirmed that the vaccine peptide had high affinity for Toll-like receptor 3 binding, and that the vaccine-receptor complex was highly stable. As our vaccine polypeptide design captures the advantages of structural epitopes and simultaneously integrates precautions to avoid relevant side effects, it is suggested to be promising for elicitation of an effective and safe immune response against SARS-CoV-2 in vivo.



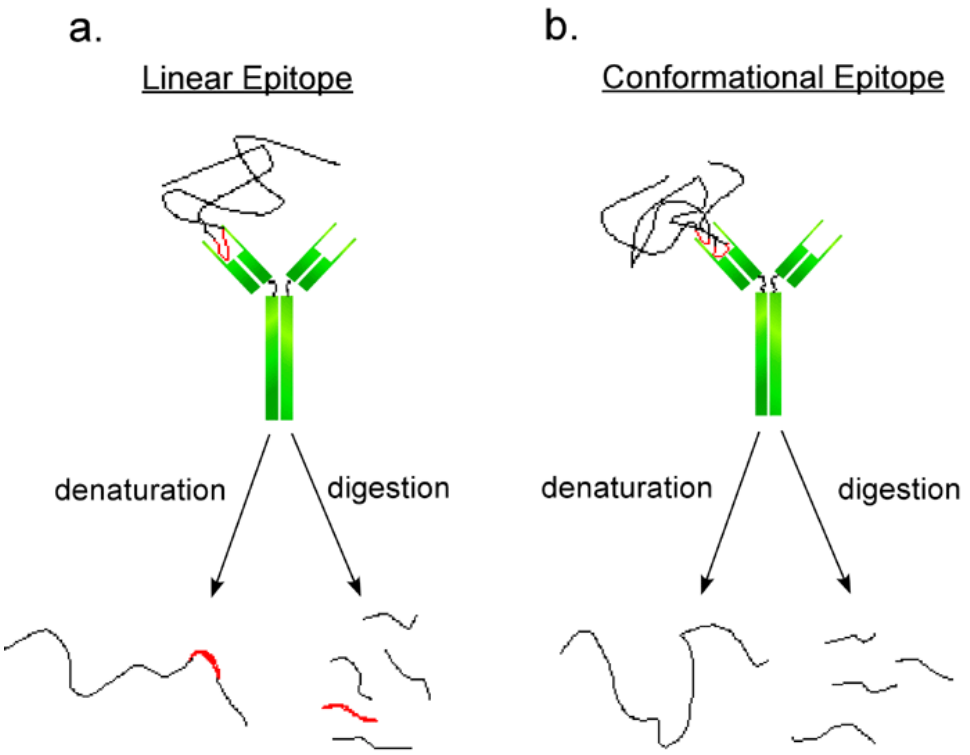
Immunogenecity :

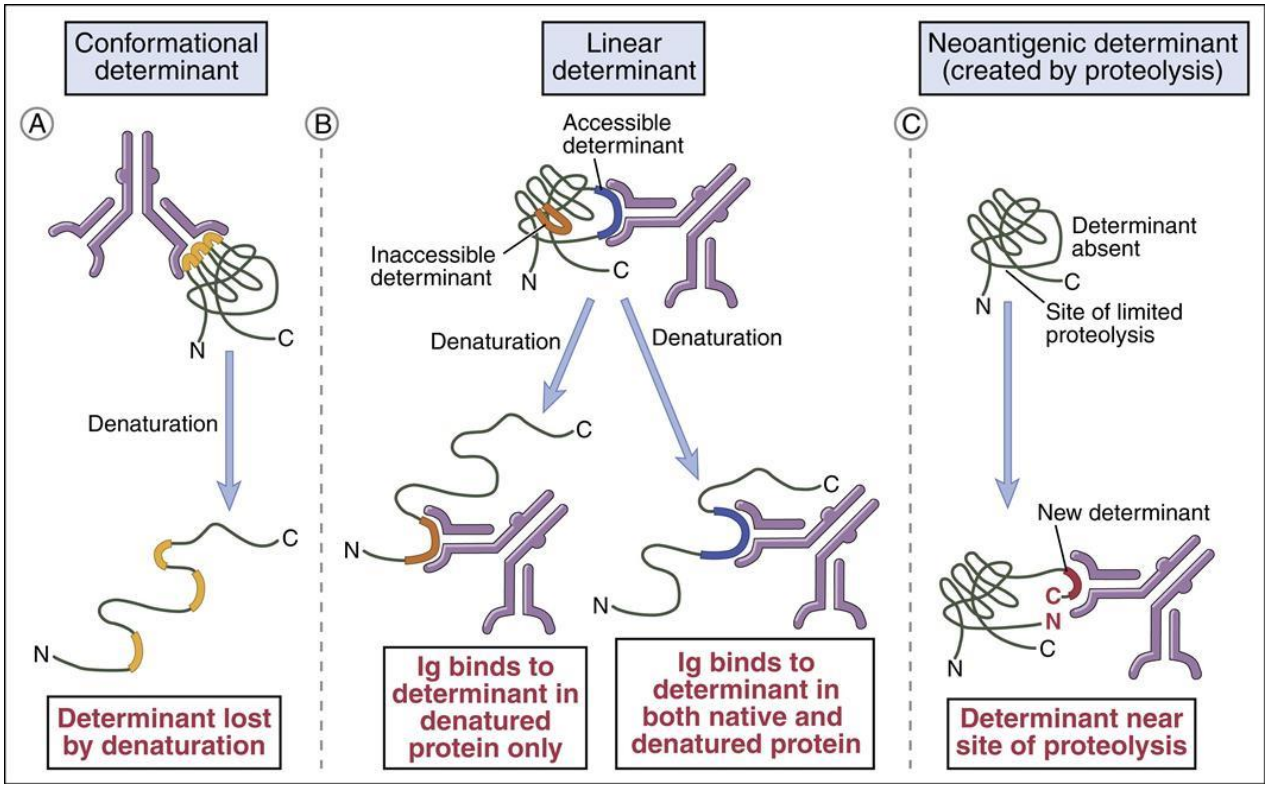
The term **immunogenicity** refers to the ability of a substance to induce cellular and humoral immune response, while **antigenicity** is the ability to be specifically recognized by the antibodies generated as a result of the immune response to the given substance.



Epitopes and Epitology

- **Antigenic determinants, or epitopes, are particular surface areas of a protein that are** specifically recognized by immunoglobulin molecules.
- Epitopes are commonly classified as either linear or conformational.
- Linear epitopes are continuous amino acid sequences of five to ten residues.
- When a protein is denatured or digested into several segments, peptides corresponding to the linear epitope amino acid sequences can sometimes be recognized by the antibody (Fig).
- Conformational epitopes are however discontinuous and occur as a result of the higher order of the structure (Fig).
- After denaturation or digestion into small fragments, the peptides corresponding to the amino acid sequences of the conformational epitopes can no longer bind the antibody (some conformational epitopes were found to be composed of a number of linear epitopes).



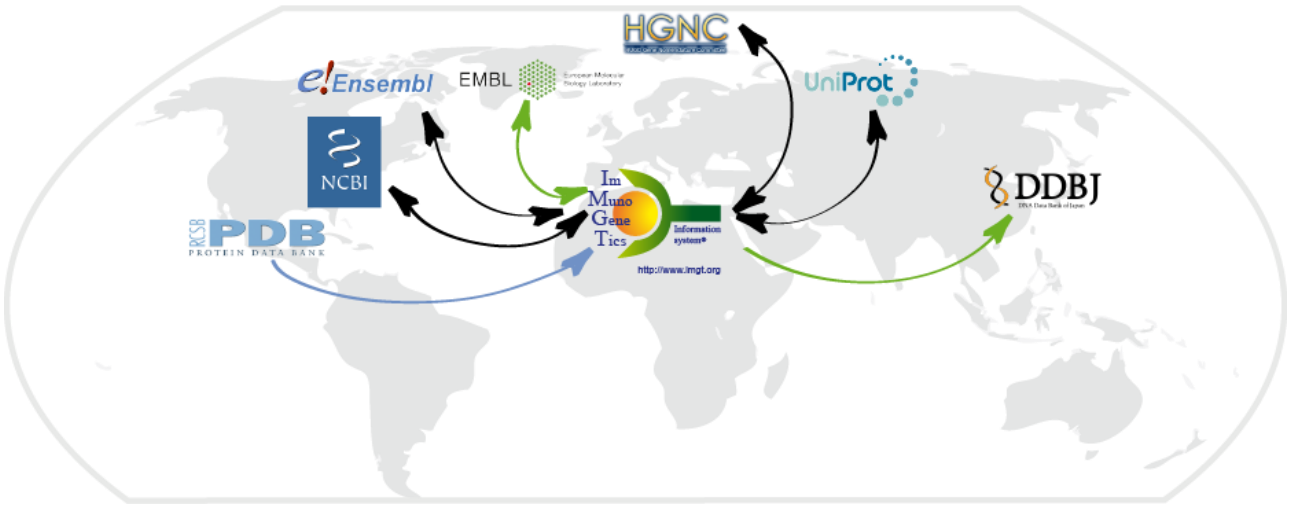


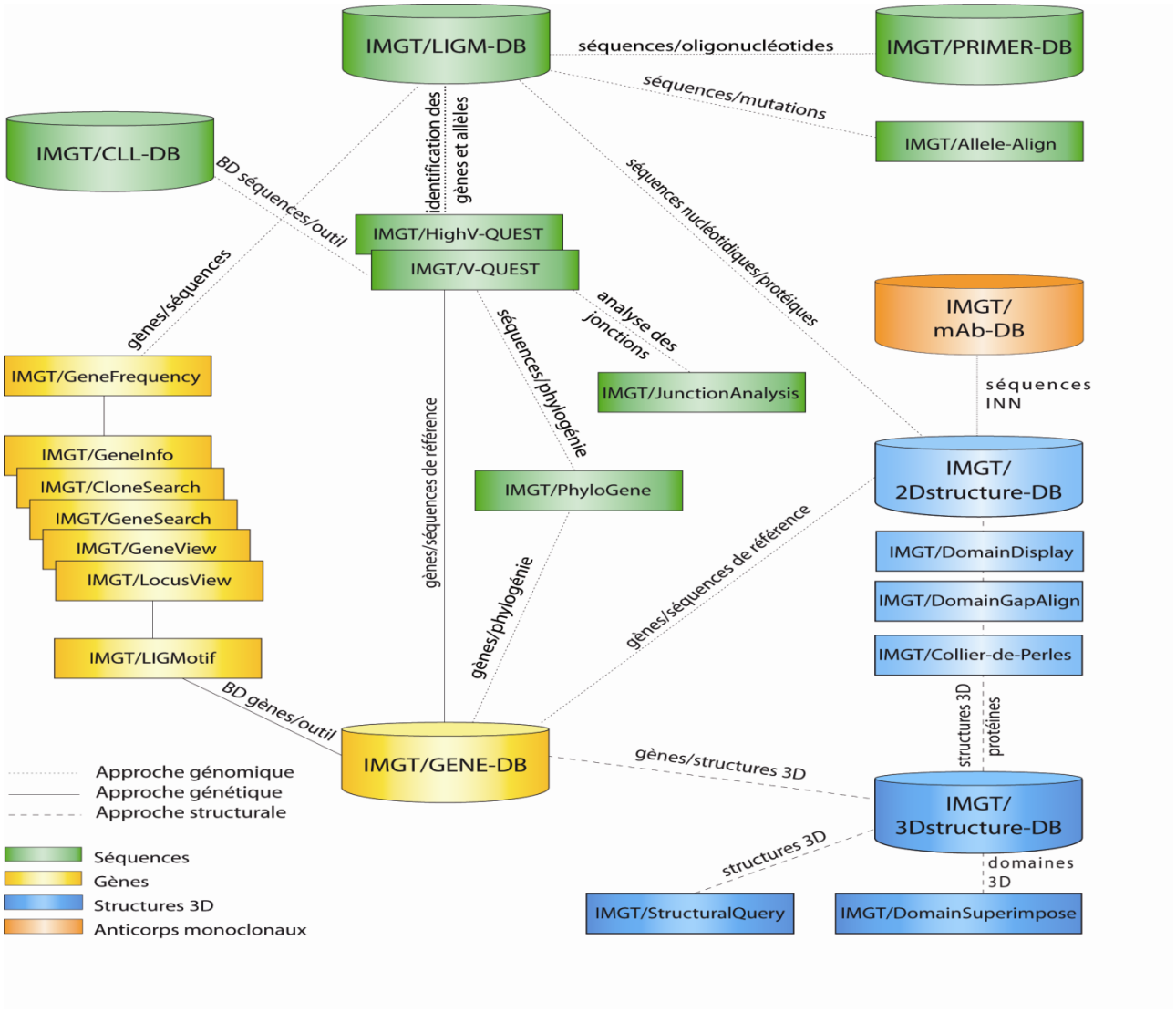
Immunoinformatics databases-

Databases	Names	URLs
B cell epitopes	CED	http://www.immunet.cn/ced/log.html
	Bcipep	http://www.imtech.res.in/raghava/bcipep
	Epiotme	http://www.rostlab.org/services/epitome/
	IEDB	http://www.immuneepitope.org/
	IMGT®	http://www.imgt.org
T cell epitopes	Syfpeithi	http://www.syfpeithi.de
	IEDB	http://www.immuneepitope.org/
	IMGT®	http://www.imgt.org
Allergen	Database of IUIS	http://www.allergen.org
	SDAP	http://www.fermi.utmb.edu/SDAP/
Information related to molecular evolution of immune system components	ImmTree	http://www.bioinf.uta.fi/ImmTree
	Immunome database	http://www.bioinf.uta.fi/Immunome/
	ImmunomeBase	http://www.bioinf.uta.fi/ImmunomeBase
	Immunome Knowledge Base	http://www.bioinf.uta.fi/IKB/

IMGT

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org, created in 1989 by Marie-Paule Lefranc in Montpellier, France (Laboratoire d'ImmunoGénétique Moléculaire (LIGM), University of Montpellier (UM) and French National Center for Scientific Research (CNRS)) has for mission to be the global reference in immunogenetics and immunoinformatics.


HGNC
e!Ensembl
EMBL
UniProt
NCBI
RCSB PDB
PROTEIN DATA BANK
Im
Muno
Gene
Tics
Information system
http://www.imgt.org
DDBJ


IMGT/LIGM-DB
séquences/oligonucléotides
IMGT/PRIMER-DB
séquences/mutations
IMGT/Allele-Align
IMGT/CLL-DB
BD séquences/outil
identification des gènes et allèles
séquences nucléotidiques/protéiques
IMGT/HighV-QUEST
IMGT/V-QUEST
IMGT/
mAb-DB
gènes/séquences
analyse des jonctions
séquences/phylogénie
IMGT/GeneFrequency
IMGT/JunctionAnalysis
séquences INN
IMGT/GeneInfo
IMGT/CloneSearch
IMGT/GeneSearch
IMGT/GeneView
IMGT/LocusView
IMGT/PhyloGene
IMGT/
2Dstructure-DB
gènes/séquences de référence
IMGT/DomainDisplay
IMGT/DomainGapAlign
IMGT/Collier-de-Perles
gènes/séquences de référence
gènes/phylogénie
IMGT/LIGMotif
BD gènes/outil
structures 3D
protéines
IMGT/GENE-DB
gènes/structures 3D
IMGT/
3Dstructure-DB
structures 3D
domaines 3D
IMGT/StructuralQuery
IMGT/DomainSuperimpose
Approche génomique
Approche génétique
Approche structurale
Séquences
Gènes
Structures 3D
Anticorps monoclonaux

Name	Description	Function	Topic
IMGT® databases	IMGT/LIGM-DB	The IMGT® database for nucleotide sequences with translation of immunoglobulins (IG) and T cell receptors (TR). Annotation is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input: ID Output: Flat file, annotation report (HTML)
	IMGT/PRIMER-DB	The IMGT® database for oligonucleotides (primers) of immunoglobulins (IG) and T cell receptors (TR). Annotation is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : ID Output : Entry cards (HTML)
	IMGT/GENE-DB	The IMGT® database for immunoglobulin (IG) and T cell receptor (TR) genes and alleles (international nomenclature). Annotation is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : ID Output : Entry cards (HTML)
	IMGT/3Dstructure-DB	The IMGT® database for 3D structures of immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) proteins, related proteins of the immune system (RPI) and fusion proteins for immune applications (FPIA). Annotation is based on the IMGT-	Query and retrieval - Input : ID Output : Entry cards (HTML)

Name	Description	Function	Topic
		ONTOLOGY concepts.	
	IMGT/2Dstructure-DB	The IMGT® database for 2D structures (IMGT Colliers de Perles) of immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) proteins and related proteins of the immune system (RPI). Annotation is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : ID Output : Entry cards (HTML)
	IMGT/mAb-DB	The IMGT® database for monoclonal antibodies (mAb) or immunoglobulins (IG), fusion proteins for immune applications (FPIA) and composite proteins for clinical applications (CPCA). IMGT/mAb-DB provides links to IMGT/2Dstructure-DB and IMGT/3Dstructure-DB.	Query and retrieval - Input : ID Output : Entry cards (HTML)
IMGT® o	IMGT/V-QUEST	IMGT® tool for nucleotide sequence alignment and analysis of immunoglobulin (IG) or antibody and T cell receptor (TR) variable domains, integrates IMGT/JunctionAnalysis, IMGT/Automat and IMGT/Collier-de-Perles.	Nucleotide sequence analysis Input : FASTA Output: Report (HTML, Excel)

Name	Description	Function	Topic
nline tools		Analysis is based on the IMGT-ONTOLOGY concepts.	
	IMGT/HighV-QUEST	IMGT® portal for NGS high-throughput nucleotide sequence analysis of immunoglobulins (IG) and T cell receptors (TR) variable domains, integrates IMGT/JunctionAnalysis and IMGT/Automat. Analysis is based on the IMGT-ONTOLOGY concepts.	Next Generation Sequencing (NGS) nucleotide sequence analysis Input : FASTA Output: Files (pdf, csv)
	IMGT/JunctionAnalysis	IMGT® tool for the analysis of the nucleotide sequences of the V-J and V-D-J junctions of the variable domains of the immunoglobulins (IG) or antibodies and T cell receptors (TR). Analysis is based on the IMGT-ONTOLOGY concepts.	Nucleotide sequence analysis Input: FASTA Output: Report (HTML)
	IMGT/Allele-Align	IMGT® tool for the comparison of two alleles highlighting the nucleotide and amino acid differences.	Nucleotide sequence alignment Input: FASTA Output: Display (HTML)
	IMGT/PhyloGene	IMGT® tool to compute and draw phylogenetic trees for immunoglobulin (IG) and T cell receptor (TR) V-REGION nucleotide sequences. Analysis is based on the IMGT-	Nucleotide sequence alignment and phylogeny trees Input: FASTA Output: Display (HTML)

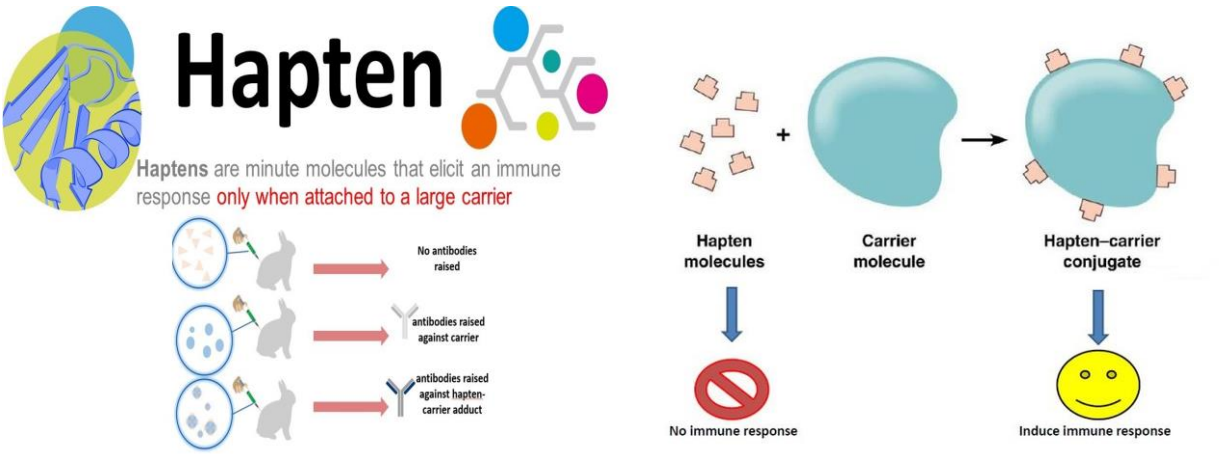
Name	Description	Function	Topic
		ONTOLOGY concepts.	
	IMGT/DomainDisplay	The IMGT® tool for the display of the amino acid sequences of the variable (V), constant (C) and groove (G) domains. Analysis is based on the IMGT-ONTOLOGY concepts.	Amino acid sequence analysis Input: FASTA Output: Display (HTML)
	IMGT/LocusView	IMGT® tool providing a view of immunoglobulin (IG), T cell receptor (TR) and major histocompatibility (MH) loci. Display is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : locus name Output: Display (HTML)
	IMGT/GeneView	IMGT® tool providing the view of a gene in IMGT/LocusView. Display is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : gene name Output: Display (HTML)
	IMGT/GeneSearch	IMGT® tool allowing the search of a gene in IMGT/LocusView. Search is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input : gene name Output: Report (HTML)
	IMGT/CloneSearch	IMGT® tool allowing the search of a clone in IMGT/LocusView. Search is based on the IMGT-ONTOLOGY	Query and retrieval - Input: clone name Output: Report (HTML)

Name	Description	Function	Topic
		concepts.	
	IMGT/GeneInfo	IMGT® tool providing combination (V-J and V-V) information for the human and mouse T cell receptor (TR) loci. Analysis is based on the IMGT-ONTOLOGY concepts.	Query and retrieval -
	IMGT/GeneFrequency	IMGT® tool providing the frequency usage of immunoglobulin (IG) and T cell receptor (TR) variable (V), diversity (D) and joining (J) genes in Homo sapiens and Mus musculus rearranged sequences from IMGT/LIGM-DB. Analysis is based on the IMGT-ONTOLOGY concepts.	Query and retrieval - Input: sequences from IMGT/LIGM-DB Output: Bar graphs of gene usage (V, D or J).
	IMGT/DomainGapAlign	IMGT® tool for the analysis of amino acid sequences of variable (V), constant (C) or groove (G) domains. Analysis is based on the IMGT-ONTOLOGY concepts.	Amino acid sequence analysis Input: FASTA Output: Display of aligned V, C or G domain sequences, gapped according to the IMGT unique numbering
	IMGT/Collier-de-Perles	IMGT® tool to draw standardized IMGT 2D graphical representations, or IMGT Colliers de Perles, of variable (V), constant (C), groove (G) domains, starting from the user own domain	Graphical representation or IMGT Colliers de Perles of V, C or G domain Input: Amino acid sequences, gapped based on the IMGT

Name	Description	Function	Topic
		amino acid sequences. IMGT Colliers de Perles are based on amino acid sequences gapped according to the IMGT-ONTOLOGY concept of IMGT unique numbering.	unique numbering Output: IMGT Collier de Perles
	IMGT/DomainSuperimpose	IMGT® tool to superimpose 3D structures of two domains from IMGT/3Dstructure-DB. Domain 3D structures are numbered according to the IMGT-ONTOLOGY concept of IMGT unique numbering.	Domain 3D structure comparison Input: 3D structure domain ID Output: Display of the superimposed 3D structures
	IMGT/StructuralQuery	IMGT® tool to retrieve IMGT/3Dstructure-DB entries using amino acid structural criteria for variable (V), constant (C) or groove (G) domains. Annotation is based on the IMGT-ONTOLOGY concepts.	Query and retrieval -

HaptenDB

- The haptens are low molecular weight molecules which by itself do not elicit immune response until and unless complexed with an immunogenic carrier, such as protein. Once an antibody is formed, it can bind to Hapten.

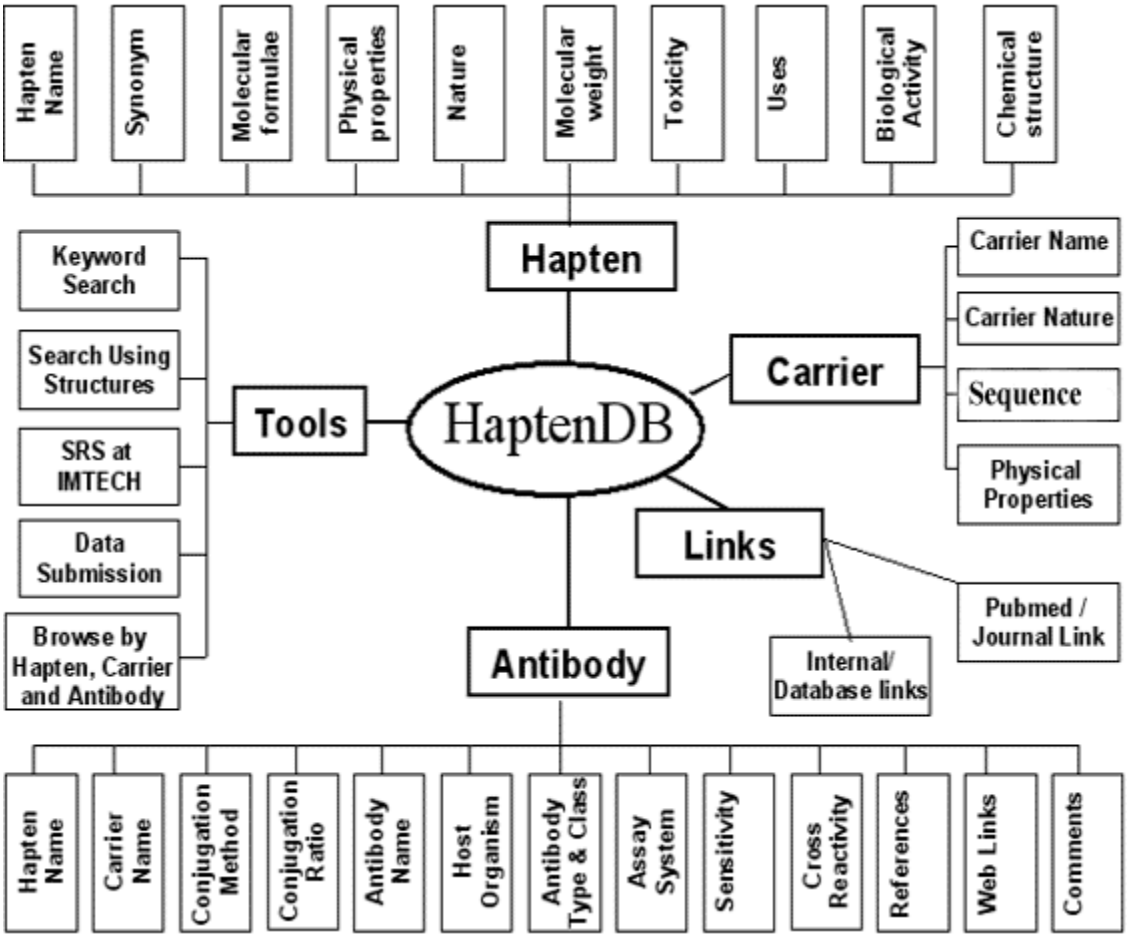


Haptendb Database

Haptendb is a database of haptens which provide comprehensive information about the Hapten molecule, ways to raise antibodies against particular group of haptens, specificity and cross reactivity of raised antibody with related haptens, use of antibodies in constructing cost effective and simple detection kits. Following are major features

- It covers wide array of haptens that includes; pesticides, herbicides, insecticides, drugs, vitamins, steroids, hormones, toxins, dyes, explosives, etc.
- The database contains 2021 entries for antibodies either raised against haptens or cross-reactivity of antibody raised against one Hapten with other related haptens. Every single record in the database contains detailed information about the Hapten, the carrier and the antibodies along with the assay methods and their sensitivity towards Hapten detection.
- The database provides information about 1087 haptens that includes: (i) common and chemical name of Hapten, (ii) molecular mass, physical and chemical properties, (iii) biological importance and the structure.
- Haptendb provides online web tools that allows users to retrieve and analyze the data that includes: (i) tools for searching database using keywords with many options and, (ii) browsing tool that allows the user to browse the database on Hapten name, carrier protein and antibody.

- The database has 2-D and 3-D structures of most of haptens in standard format based on information in literature. It also allows sketching structures online and searching of similar structures in database.
- One of the powerful tools in Haptendb is structure similarity search tool, which allows user to search similar structures.



Hapten Details	
Hapten_Name / Synonyms / Modification	Chemical name of Haptenic Compound, its common name or modification in an existing well known compound by introducing some groups or replacing one group with other.
Hapten_Nature	Nature or Category of the haptenic compounds e.g. Pesticide, Drug , Peptide, Hormone, Vitamins etc.
Formula / Weight / Physical	These fields include emperical formula of hapten, molecular weight

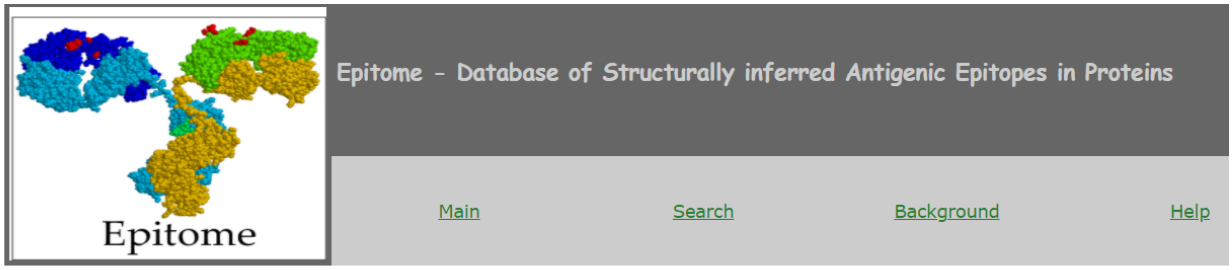
Properties	and other physical properties such as Melting point, Boiling Point, Density etc.
Uses / Biological Activity	These fields contain information about the different uses of the hapten and their actions and affects in biological systems
Carrier Details	
Carrier_Name	Name of the Carrier Macromolecule that is used for generating immunogenic conjugates by conjugation with the hapten.
Carrier_Nature / Chemical & Physical Properties	Nature or Category of the Carrier Macromolecule . They are usually larger protein molecules such as Bovine Serum Albumin (BSA), Ovalbumin (OVA or OA), Keyhole Limpet Heamocyanin (KLH) etc.
Antibody Details	
Conjugation_Method & Method_Detail	The method used for the conjugation of hapten with the carrier molecules. They are well defined protocols that are usually used with some modifications and are cited in litrature e.g. Active Ester Method, Mixed Anhydride Method.
Spacer / Linkage_Nature	The spacer arm that is attached to hapten before conjugation to carrier molecules. Linkage Nature the nature of bond between the hapten and the carrier molecule e.g. Amide linkage
Hapten_Carrier_Ratio	Number of haptens attached per molecule of carrier.
Antibody_Name	Name of the Antibody that is raised against hapten by immunizing some organism (e.g. mice, rabbit, sheep etc.)
Type & Class	It is the Type of Antibody that are raised in the host organism e.g. Monoclonal, Polyclonal or only antiserum. Furhter if Monoclonal Antibodies are raised then furthrer to which type belongs too e.g. IgG, IgM etc.
Specificity	Specificity of the Raised antibodies for the target haptenic compound
Cross-reactivity	Cross-reactivity of the raised antibodies with other similar or related compounds. This is either expressed in percentage (which is expressed as IC50 of the tested compound x 100 / IC50 of target

	compound where IC50 is referred as amount required for 50% inhibiton of the antibody in the given set of conditions) or directly as IC50 value.
Sensitivity	This is also referred as limit of detection of the hapten with the raised antibody.
Assay_System	The method used for charactizing the antibodies e.g. competitive ELISA, non-competitive ELISA, RIA etc.
Application	Likely application and future prospects of the ELISA method developed, Antibody raised etc.
Reference	This field has the detail of the Journal , Author, Title, Volume, Page Nos., Year of Publication of the paper in which this information is reported.
Web_link	This field contain the web link of the research paper that is cited in the refrence field.
Comments	This field contain other relevent information that is not contained in all the above mentioned fields such as immunization protocol, some other important properties of antibody, hapten or carrier.

EPITOME

To compile this database, we aligned all available structures of antibodies and analyzed them to identify CDRs. Based on this analysis we found all the proteins in PDB that are bound to an antibody, and identified within them the residues that bind to CDRs.

Each entry in the database describes one interaction between a residues on an antigenic protein and a residues on an antibody chain. Every interaction is described using the following parameters: (1) PDB ID (2) PDB chain ID of the antigenic protein (3) PDB position of the antigenic residue (4) type of antigenic residue and its sequence environment (5) PDB chain ID of the antibody chain (6) type of antibody chain (heavy or light) (7) CDR number (8) PDB position of the antibody residue (9) type of antibody residue and its sequence environment.



Epitome is a database of all known antigenic residues and the antibodies that interact with them, including a detailed description of residues involved in the interaction and their sequence / structure environments. Additionally, Interactions can be visualized using an interface into Jmol.

Publication

Complementary resources

- Dr. Andrew C.R. Martin's Group at UCL - general information about antigens and antibodies http://www.bioinf.org.uk
- The international ImMunoGeneTics information system http://imgt.cines.fr/
- Darren Flower 's databases of quantitative functional peptide data for immunology:
    1. JenPep - http://www.jenner.ac.uk/JenPep
    2. AntiJen - http://www.jenner.ac.uk/antijen/
- BciPep - a database of B cell epitopes http://www.imtech.res.in/raghava/bcipep
- The HIV Molecular Immunology Database http://hiv-web.lanl.gov/content/immunology/index.html

dbMHC

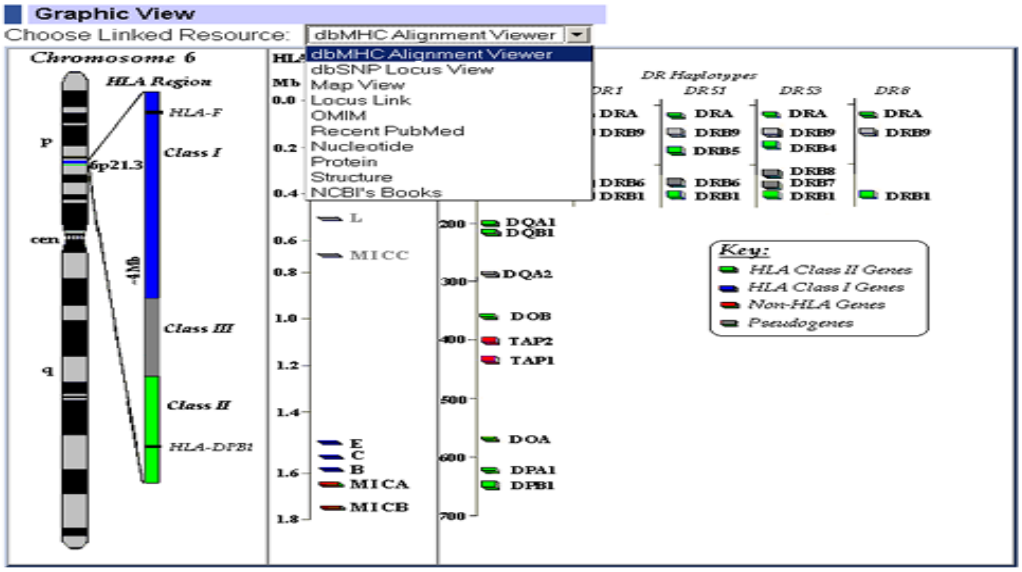
Major Histocompatibility Complex database (dbMHC)

- The dbMHC database provides an open, publicly accessible platform for DNA, and clinical data related to the human Major Histocompatibility Complex (MHC).
- The need to share research and clinical data focused on the MHC has lead to a series of meetings at the International HLA WorkShop & Congress (IHWC).
- The data generated from the 13th IHWC is presented at NCBI in dbMHC.
- In addition, the dbMHC will provide tools for submission and analysis of research data linked to the MHC.

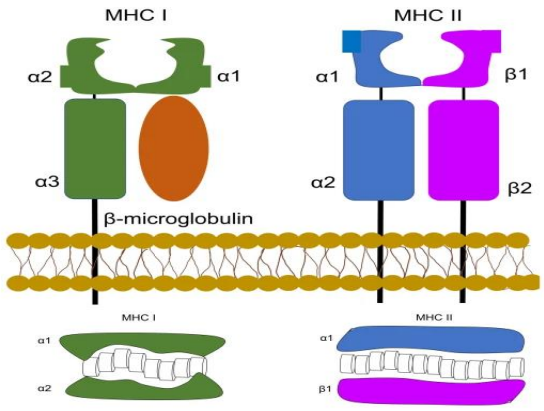
What is MHC and why does it matter?

- The major histocompatibility complex (MHC) genes code for proteins which the immune system uses to identify cells and tissues in the body as “self” or “other”.
- MHC molecules ‘talk’ to T cells which patrol the body for foreign invaders or dangerously mutated cells.
- The MHC acts as a window into our cells.
- It presents snippets of information (peptides) on the state of the cell- allowing the immune system to check for infection, cancer, and other maladies.
- Cells that do not pass the self/other test are eliminated.

- The MHC genes are polymorphic, and individual organisms from the same species very rarely have the same MHC identity. When looking for a donor for organ or tissue transplantation, both the donor and recipient MHC identities are determined, to find the best and closest match. Your MHC identity is also called your "tissue type". This tissue type is critical in organ transplantation- mismatches make grafted or transplanted tissues a target for the adaptive immune system.
- MHC molecules are comprised of two individual parts that present short epitopes (short peptides) to cells of the immune system. There are two main classes of MHC molecule – Class I and Class II. There are also "non-canonical" MHC types which serve specialized purposes and present specialized molecules.
-



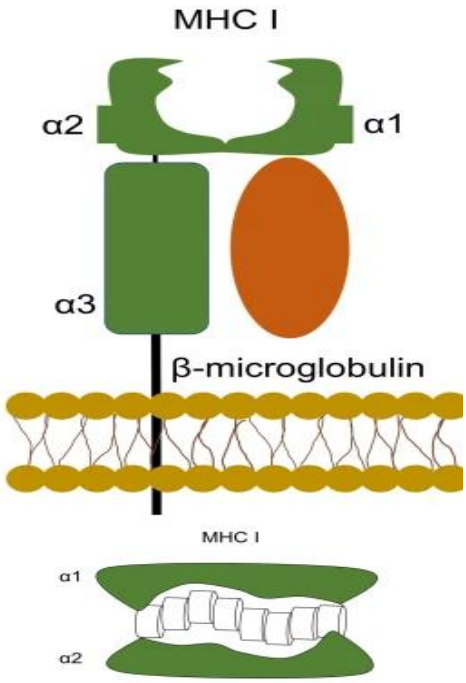
-



The Classical MHC molecules
MHC Class I and Class II present peptides to immune cells as part of routine immune surveillance.
LEFT: MHC Class I is comprised of three alpha subunits and beta macroglobulin. The binding groove (lower left) of Class 1 is deep, with closed ends and binds peptides of 8-10 amino acids in length.
RIGHT: MHC Class II is a heterodimer of a 2-unit alpha chain and a 2-unit beta chain. The binding groove of class II is shallow and open at each end, allowing binding of peptides 13-17 amino acids in length.

-
-

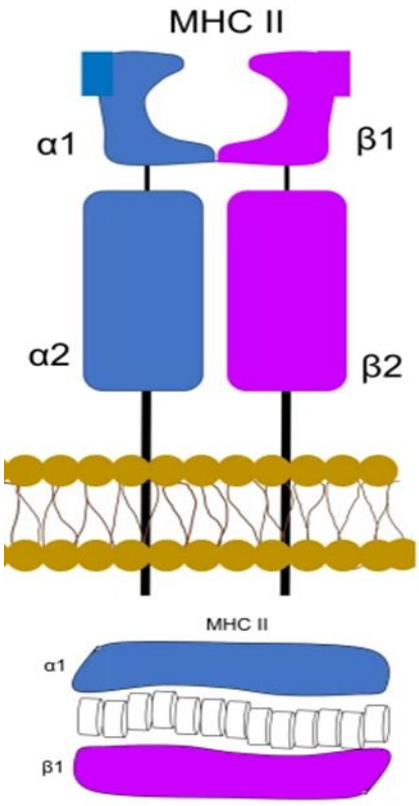
- Class I MHC molecules are found on all nucleated cells in the body and on platelets.
- Class I interacts with CD8+ T cells, interacting directly with CD8 as a co-receptor.
- Presentation of intracellular epitopes allows T cells to check for intracellular bacteria, viral infection and cancerous mutations. MHC I presentation and signaling is a global “alarm” system for the cells in the body.
- The Class I MHC molecule is made of 2 proteins- a three-domain alpha unit non-covalently bonded to beta-2 microglobulin.
- The amino acid sequence and shape of these subunits determines the shape of the binding groove and therefore what peptide can bind.
- MHC Class I present epitopes of 8-10 amino acids to T cells, typically derived from proteins in the cytosol (endogenous protein antigens).



-
- Class II MHC molecules are typically found on antigen presenting cells (APC) such as macrophages, dendritic cells, and B lymphocytes. These MHC molecules interact with CD4 on CD4+ T helper cells. Class II MHC presentation functions as a specific line of communication between immune cells and the global immune system.
- The Class II molecule is made up of two distinctive subunits- alpha and beta, which are non-covalently linked to form the binding groove. In that groove, epitopes derived from extracellular contents are presented in 14-18 amino acids peptides. Class II MHC

presentation is a requirement for initiating and sustaining adaptive immune responses against foreign invaders such as fungi and extracellular bacteria.

- This same family of proteins in humans are called Human Leukocyte Antigens (HLA).



The dbMHC is divided into two main sections,

a Reagent Database section and a Clinical section.

The Reagent database contains the reagent data needed to trace DNA typing. This section provides an open platform for the submission, evaluation, and editing of individual reagent specifications of Sequence Specific Oligonucleotides and Sequence Specific Primers as well as typing kit information. All reagents are characterized for allele specificity using the current curated World Health Organization HLA allele database in cooperation with IMGT/HLA.

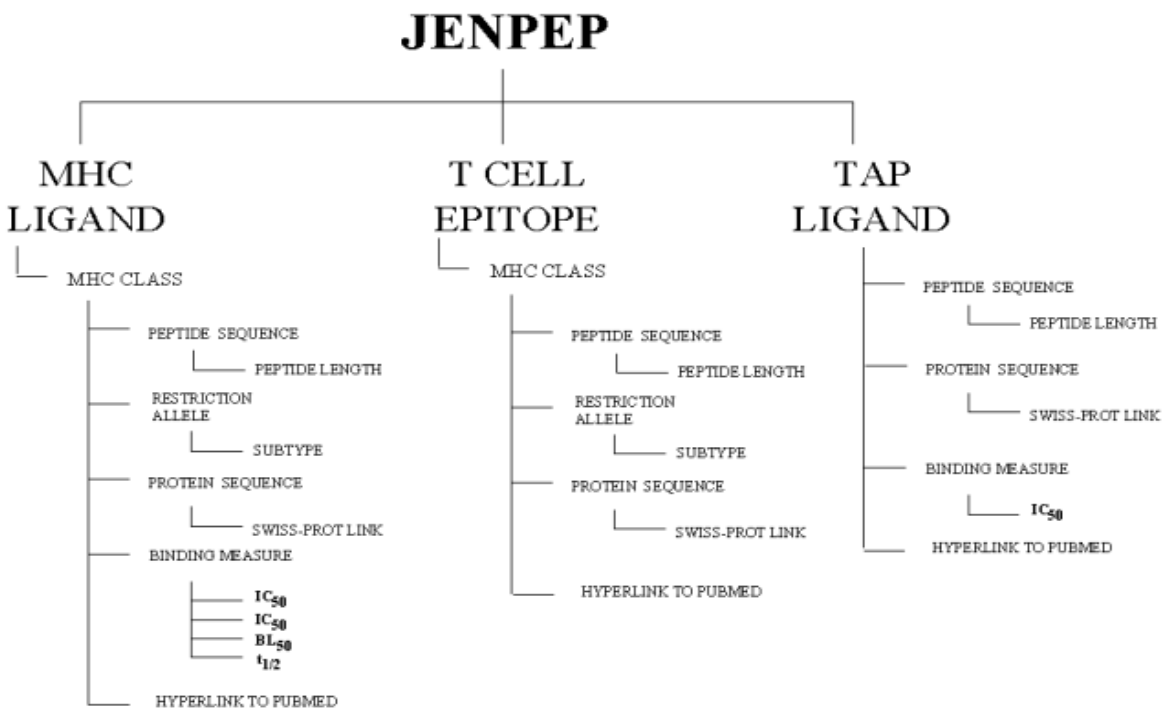
The Clinical section will contain anonymous clinical data from individuals taking part in MHC-related research projects in the general categories of Anthropology, Cytokine Polymorphisms, HLA-E,F,G, Cancer, Disease, HLA Alloantibodies & Kidney Graft Rejection, Mycobacterial Disease, New Allele Registry, Hemochromatosis/ Psoriasis, and Virtual DNA Analysis

JenPep

Motivation: The compilation of quantitative binding data underlies attempts to derive tools for the accurate prediction of epitopes in cellular immunology and is part of our concerted goal to develop practical computational vaccinology.

Results: JenPep is a family of relational databases supporting the growing community of immunoinformaticians. It contains quantitative data on peptide binding to Major Histocompatibility Complexes (MHCs) and to Transmembrane Peptide Transporter (TAP), as well as an annotated list of T-cell epitopes.

Availability: The database is available via the Internet. An HTML interface allowing searching of the database can be found at the following address: http://www.jenner.ac.uk/JenPep.



Epijen

EpiJen is a reliable multi-step algorithm for T cell epitope prediction, which belongs to the next generation of *in silico* T cell epitope identification methods. These methods aim to reduce subsequent experimental work by improving the success rate of epitope prediction.

EpiJen step one: proteasome cleavage

The dataflow in EpiJen is shown in Figure Initially, the protein is chopped into overlapping decamers and processed by a proteasome cleavage QM. A previously derived and tested p1p1' model, as described in the Methods section below , is used. The model takes into account only the contributions of the residues next to the cleavage site: C-terminus and the next aa. Two thresholds, 0.0 and 0.1, can be used here. Threshold 0.0 is recommended for alleles which prefer Phe or Trp at the C-terminus: HLA-A*24, HLA-B*07, HLA-B*27, HLA-B*35, HLA-B*51 and HLA-B*53. The epitopes for other alleles are predicted accurately at a threshold of 0.1. This initial step has a powerful filtering ability: between one half and two thirds of the true negatives were eliminated by this step. The "cleaved" peptides, present as nonamers, are then passed to the next filter: the TAP binding QM.

EpiJen step two: TAP transport

The TAP binding QM also has been derived and tested previously.A threshold of 5.00 is recommended for both fully and partially TAP-dependent alleles. Pro and Asp at anchor position 2 has a strong negative effect on TAP binding . For that reason, a threshold of 3.0 is recommended for epitopes binding to HLA-B*07, HLA-B*35, HLA-B*40, HLA-B*44, HLA-B*51 and HLA-B*53. The filtering ability of the TAP step is low. Up to 10% of the true negatives are eliminated here. The "transported" peptides move to the next filter: MHC binding.

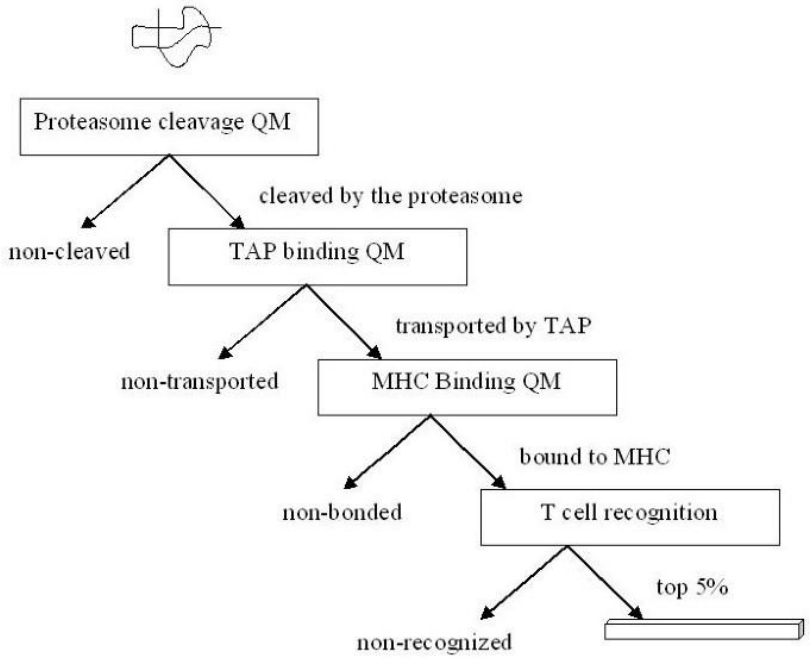
EpiJen step three: MHC binding

EpiJen includes 18 QMs which can be used to predict binding to different HLA-A and B alleles. Certain QMs were developed for single alleles and others developed for allele families. QMs developed for whole supertypes were poorly predictive, especially for HLA-B supertypes. Some MHC models were derived previously, while others were developed for this study. Quantitative data (continuous values like IC_{50}s) were available for certain alleles, for the rest only sequences of binders were known (discontinuous values). As is described in the Methods section below, binding models based on continuous values were derived by multiple linear regression (MLR) (Table 1) and those based on discontinuous values by discriminant analysis (DA) (Table 2). "Leave-one-out" cross-validation tests indicate a higher predictive rate for the DA models ($AUC_{ROC} > 0.9$; *accuracy* > 80%) than MLR models ($q^2 \approx 0.5$). The filtering ability of this step is significant:

approximately 25–30% of the true negatives are eliminated here. The thresholds for this step are 0.5 for the DA models and 5.3 for MLR models. These thresholds can not be altered by the user. They seek to reduce the number of false positives in long

EpiJen step four: epitope selection

All peptides which are presented by MHCs on the cell surface after being cleaved by the proteasome and transported by TAP could potentially be T cell epitopes. However, only a small number of all possible epitopes are actually immunogenic. To reduce the number of false positives we tested different thresholds, which we defined as percentages of available peptides sourced by one protein. The top 5% threshold performed best, giving 85% sensitivity; we recommend it and use it as a default value for this step. Optional are thresholds 2, 3 and 4%.



Monoclonal antibodies database.

Antibody Related Databases and Software

Antibody related amino acid sequencing tools, nucleotide sequencing tools, structural modeling tools, and hybridoma/cell culture databases can be found below. Speciality research databases that include monoclonal and polyclonal antibodies are also included.

- ABG: Directory of 3D structures of antibodies - The directory, created by the Antibody Group (ABG), allows a quick access to the antibody structures compiled at the Protein Data Bank (PDB). In the directory, each PDB entry has a hyperlink to the original source to make full information recovering easy. The VH and VL sequences were aligned and are reported in the VH and VL alignments. The directory will be updated monthly.

- **ABG: Germline gene directories of the mouse** - a directory of mouse VH and VK germline segments, part of the webpage of the Antibody Group at the Instituto de Biotecnologia, UNAM (National University of Mexico)
- **AbMiner** - users can search for commercially available antibodies and match them with their respective genomic identifiers. A unique feature of this database is that only antibodies screened by Western blot are included, which makes the antibodies listed research friendly.
- **Abnum** - an online tool that lets you number an antibody sequence or structure automatically using the Kabat or Chothia schemes, as well as a new improved Clothia scheme created by the authors.
- **Antibody Central** - an antibody search portal which connects catalog antibodies and research publications to the UniProt protein information resource. The tool uses a proprietary Antibody Name Dictionary model - a collection of protein names, synonyms, and symbols recognized as antibody names which are verified through a full-text search of scientific publications in HighWire Press.
- **bNAber.org** (database of HIV-1 broadly neutralizing antibodies) to the list of "Antibody Related Databases and Software
- **European Collection of Cell Cultures** - a cell culture collection service for the research community that holds over 40000 cell lines including 450+ antibodies
- **HIV Molecular Immunology Database** - listings of HIV specific monoclonal antibodies, maps of binding sites on HIV proteins, sequence alignments showing global variation of linear binding domains, brief descriptions of how the MAb has been used in HIV studies with Medline links, and where to obtain many of the antibodies in the database.

- **The Hybridoma Databank** - HDB holds data on various aspects of hybridomas and their immunoreactive products. Information on a hybridoma's construction and the reactivity and non-reactivity of its secreted product is included. In addition, information on the availability of an individual hybridoma and its Mab product are included. Information in the HDB is derived from literature, catalogs and survey forms.
- **IMGT ®, the international ImMunoGeneTics information system ®** - created in 1989 by Marie-Paule Lefranc (Université Montpellier II, CNRS), IMGT is an integrated knowledge resource specializing in immunoglobulins, T cell receptors, major histocompatibility complex, immunoglobulin superfamily , major histocompatibility complex superfamily, and related proteins of the immune system for human and other vertebrate species. IMGT consists of sequence databases (IMGT/LIGM-DB, a comprehensive database of IG and TR from human and other vertebrates, with translation for fully annotated sequences, IMGT/MHC-DB, IMGT/PRIMER-DB), a genome database (IMGT/GENE-DB), a structure database (IMGT/3Dstructure-DB), a web resource (IMGT Repertoire), and interactive tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, etc...). The IMGT home page provides a common access to all Immunogenetics data.
- **The Kabat database** - search the Kabat database of sequences of proteins of immunological interest. This site is updated frequently.
- **Kabatman** - A great source of information for molecular biologists and crystallographers on the structure of antibodies. This site also contains the KabatMan database which allows you to search the Kabat sequence database to look for sequence features, AbCheck which lets you test your antibody sequence against the Kabat sequence database to look for unusual features, and Chothia which lets you find canonical assignments for the CDRs in your antibody sequence.
- **Macromoltek** - Molecular simulations simplified. Antibody modeling, side-chain packing, renumbering, and other web-based computational tools for antibody development available in easy-to-use workspaces. Macromoltek also provides a consulting service for customers in need of more specialized analysis. Test our features with a free trial!

- **Monoclonal Antibody Index** - a biotechnology database with yearly updated information on more than 9000 monoclonal antibodies produced for the diagnosis an-d therapy of cancer, transplant, infection, heart-related disorders, etc...
- **PIGS** - **P**rediction of **I**mmuno**G**lobulin **S**tructures - use this web interface to automatically model immunoglobulin variable domains based on the canonical structure method. The output is a 3D model of the target antibody that can then be downloaded or displayed online.
- **RosettaAntibody** - a FV homology modeling server - a homology modeling server that predicts antibody Fv region structures using knowledge-based techniques for template selection, *de novo* loop modeling for creating CDR-H3, grafting for non-H3 CDR loops, and docking to optimize the orientation of the light and heavy chains.
- **SACS - (Summary of Antibody Crystal Structures)** is an automatically updated summary of all antibody structures in the Protein Data Bank (PDB). New PDB files processed immediately as they arrive to determine whether they are antibodies and if so, relevant information is extracted and made available on the web.
- **Scaligner** - Scaligner is a platform for analyzing antibody sequences: CDR identification, amino acids numbering using Kabat and IMGT schemes, multiple sequences alignment, and storage in a secure repository.
- SCOP - search "SCOP" (structural classification of proteins) for references and structural information on antibodies by keyword or sequence
- **STATdxPathIQ** - formerly Immunoquery. STATdxPathIQ is a diagnostic decision making tool for pathology that uses current, peer-reviewed literature to suggest immunohistochemical panels for the differentiation of similarly appearing tumor types. The website also contains high quality typical and atypical histochemical and immunohistochemical images to aid researchers in differential diagnosis. STATdxPathIQ evolved from ImmunoQuery, a meta-analysis search engine for use by pathologists.
- **V BASE** - V BASE is a comprehensive directory of all human germline variable region sequences compiled from over a thousand published sequences, including those in the current releases of the Genbank and EMBL data libraries.

- **WAM** - Web Antibody Modelling - WAM is an online facility for generating a 3D model of an antibody Fv from its sequence. Based on Oxford Molecular's AbM, WAM uses sequence homology to build the frameworks, the most sequence-homologous member of the canonical class to construct the canonical loops, and a combination of database and conformational searches to construct the non-canonical loops. The putative models are then screened using a combination of energetic and knowledge-based approaches. Modelling is free for the first model for academic users; there is a small charge ($100) for subsequent models. See the web site for more details for commercial users.

UNIT – II -IMMUNOINFORMATICS&COMPUTATIONAL VACCINOLOGY-SBIA5202

UNIT II GENOTYPING METHODS AND DISEASE ASSOCIATION

Genotyping of SNPs

SNP genotyping is the measurement of genetic variations of single nucleotide polymorphisms (**SNPs**) between members of a species. It is a form of **genotyping**, which is the measurement of more general genetic variation. **SNPs** are one of the most common types of genetic variation.

SNPs (single nucleotide polymorphisms) or point mutations are the most common types of genetic variation determining to a major part the phenotype diversity between individuals. Causal point mutations change the amino acid sequence of the encoded protein and hence such SNPs are involved in the characteristics of an individual.

For genome-wide SNP genotyping, e.g. population studies, association studies, genomic selection including the analysis of 3k-3000k we recommend and offer Illumina BeadChips, Affymetrix GeneChips or Next Generation Sequencing.

For gene-wide SNP genotyping, e.g. fine mapping or haplotyping of candidate regions including the analysis of <500 SNPs we recommed Fluidigm Biomark, NGS or array-based technologies.

For the analysis of individual SNPs, e.g. determination of specific mutations in phamacogenetics or diagnostics with a small number of SNPs included, we offer Realtime PCR assays, Sanger Sequencing and the above mentioned technologies.

Genotyping technologies

Real-time PCR for genotyping

Genotyping by real-time PCR is a rapid, reliable approach widely used for the confirmation of SNPs and CNVs and to small numbers of markers in hundreds or even millions of samples.

Genotyping microarrays

Used in many of the world’s largest genetic studies, the innovative Axiom Genotyping Solution is a portfolio of array-based tools ideal for everything from genome-wide analysis to routine screening of complex genetic traits.

Targeted genotyping by sequencing

Fast, affordable next-generation sequencing platforms help you discover and analyze genetic variants using genomic and targeted sequencing methods.

Digital PCR for genotyping

Digital PCR uses molecular counting to provide highly precise, sensitive results, and is ideal for low-frequency/rare allele detection.

Fragment analysis by capillary electrophoresis

Your genetic analyzer is more powerful than you may think. One platform performs both sequencing and a multitude of fragment analysis applications including SNP genotyping.

Biobanking

We offer the most diverse, cutting-edge portfolio of biobanking products and services.

Classical HLA typing

What is HLA?

- HLA stands for Human Leukocyte Antigen: a molecule that is present on almost every cell type in the human body.
- HLA has a central role in coordinating immune responses through a variety of different mechanisms.
- The main function of HLA is to present small protein molecules (‘peptides’) to T cells, a type of immune cell.
- Peptides may be from inside the cell, or from within the cell’s environment.
- If the cell is infected with a microorganism such as a virus, HLA will present viral peptides to T cells and trigger an immune response (which includes the infected cell being killed).

- Similarly, the cell may absorb molecules or microorganisms from the environment and, when presenting these to the T cell, cause the T cell to become activated and to begin searching for the source of any ‘non-self’ molecules.
- Furthermore, because HLA is present on almost all cells, immune cells check for the presence of HLA and can kill cells where HLA is absent as this indicates the cell isn’t functioning correctly (it may be a tumour cell, or infected with virus).

What is Tissue Typing?

- HLA is actually a family of molecules encoded by the HLA genes.
- ‘Classical’ HLA molecules are typically of interest clinically and in most research environments, and consist of the ‘Class I’ molecules HLA-A, HLA-B and HLA-C, and the ‘Class II’ molecules HLA-DR, HLA-DQ and HLA-DP.
- The Class I genes *HLA-A, HLA-B* and *HLA-C* encode the alpha chain of the corresponding molecule.
- HLA Class II is divided into alpha and beta chains encoded by separate genes, for example, *HLA-DQA1* and *HLA-DQB1*.
- All individuals possess the *HLA-DRB1* gene, but many individuals also possess additional *HLA-DRB3, HLA-DRB4* and *HLA-DRB5* genes.
- Class I and Class II HLA molecules have distinct functions and are present on different cell types.
- HLA genes are highly polymorphic.
- This means that, unlike most genes in the body, there are many different variants of the HLA genes (‘alleles’).
- Because of this, unrelated individuals are very likely to have different HLA alleles.
- The process of identifying which HLA variants an individual possesses is colloquially known as Tissue Typing or, more accurately, HLA typing.

- Depending on the method used, the terms HLA genotyping or serological HLA typing may be used.
- HLA genotyping uses ‘molecular’ or DNA-based methods and is considered more reliable.

Why is HLA Important?

- In the natural world, the variation seen in HLA genes means that certain variants provide more protection against infectious diseases than others.
- Differences in HLA between individuals means that, as a population, survival is more likely.
- HLA is critically important in transplantation.
- Genetic differences between the donor and recipient can illicit an immune response and, as the HLA molecule is central in the immune response pathway, differences in HLA can have a dramatic impact on transplant outcome.
- Reactivity against donor HLA can develop post-transplant, leading to rejection of the transplant by the recipient’s immune system.
- Pre-transplant exposure to non-self HLA (e.g. via blood transfusion) can also cause antibodies to develop which can result in immediate loss of the transplant. HLA is important in both the solid organ (e.g. kidney) and blood and marrow transplant (BMT) settings.
- In BMT, transplanted donor cells mature into immune cells and can attack the recipient’s own cells unless a high-level HLA match is achieved.
- Specific HLA types are also associated with an increased susceptibility to disease, or reactivity against particular drugs.
- For example, HLA has been linked to Coeliac Disease, Rheumatoid Arthritis and Narcolepsy, and HLA-B*57:01 is associated with hypersensitivity to Abacavir (an antiretroviral drug). Often the exact mechanism linking HLA to disease or drug hypersensivity is unknown.

How is HLA Typing Performed?

- Historically, multiple methods have been used for HLA typing and many different methodologies are still in use today.
- Serological HLA typing has fallen out of favour as this method uses biological reagents which are often in limited supply, require rigorous QC and provide limited discrimination between different HLA variants.
- Typically, laboratories perform HLA genotyping using molecular (DNA-based) methods which differ in terms of their level of resolution.
- For example, VH Bio Ltd. performs intermediate-resolution genotyping by PCR-RSSO and high-resolution genotyping by Next Generation Sequencing (NGS).
- High-resolution genotyping tends to define HLA alleles unambiguously (e.g. HLA-A*02:01) or to a level considered functionally important, whereas intermediate-resolution genotyping defines group of related alleles (e.g. HLA-A*02:01/02:04/02:17).
- Different methodologies have pros and cons, e.g. cost, throughput and turnaround time, level of resolution.
- VH Bio Ltd. provide a bespoke service and endeavour to act as a laboratory partner: supporting decision-making, providing HLA genotyping services, and assisting in the understanding of HLA genotyping data.

Why Perform HLA Typing?

- Clinical laboratories typically perform HLA typing to support solid organ transplantation, blood and marrow transplantation, and transfusion (effectively, a blood transplant). Typically, laboratories look to match donor and recipient HLA as closely as possible and avoid mismatches to which there are antibodies resulting from previous exposure. This field of clinical science is often referred to as Histocompatibility and Immunogenetics (H&I).
- Studies may perform HLA typing to determine whether specific HLA types are associated with susceptibility or protection to specific diseases or drug reactions. For example, early in the SARS-CoV-2 pandemic research groups sought to identify whether specific HLA types were associated with or protective against severe COVID-19 disease.

- Different HLA variants bind peptides differentially (this is referred to as the 'peptide repertoire'). Because of this, characterised and consistent HLA types are important in studies of the immune response as differences can impact the magnitude and outcome. In particular, studies involving T cells and the intimate interaction between HLA and the T cell Receptor (TCR) require careful consideration of the HLA molecule.
- Healthcare is also moving in the direction of personalised medicine. Cell-based therapies require consideration of HLA as incompatibility with the recipient can result in failure of the therapy or unwanted side effects. Cell and tissue banking services often HLA type their inventory for this reason.

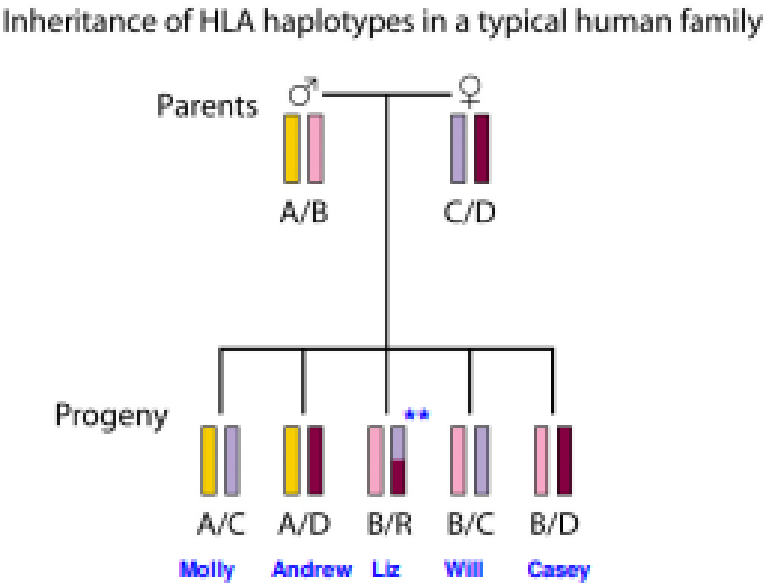
Method	About method	Pros	Cons
Serotyping	Non-sequencing based typing method where antibodies specific to HLA proteins are used to identify the proteins on the cell surface.	- Low Cost - Rapid - Tradition	- Crude Method - Protein based detection - Inaccurate typing - Protein binding to more than one serotype
Sequence Specific Oligonucleotide Hybridization (SSO)	Typing .method where specific oligos are first designed for genes of interest and then hybridized to patient or donor DNA to check for hybridization.	- Checking of specific target - Efficient	- Cannot account for unrecorded alleles - Hybridization errors - Need to know target sequence - Cannot phase
Sanger Sequencing	Sanger sequencing or Sequencing by Termination (SBT) is a classical method used for sequencing specific regions of the MHC.	- Used to sequence regions of interest target - Fast - Base pair resolution - Coverage only 2X	- Different HLA alleles share similar sequences, difficulty aligning - Cannot phase
Next-gen Sequencing	Performing long range PCR to amplify HLA genes in MHC region, fragmenting the amplified genes.	- Deep coverage (1000x) - Total MHC coverage - Rapid high throughput - Accurate and efficient - Phasing	- Data Analysis

MHC haplotypes

Terminology: • Haplotype: set of alleles present in each parental chromosome (two sets).

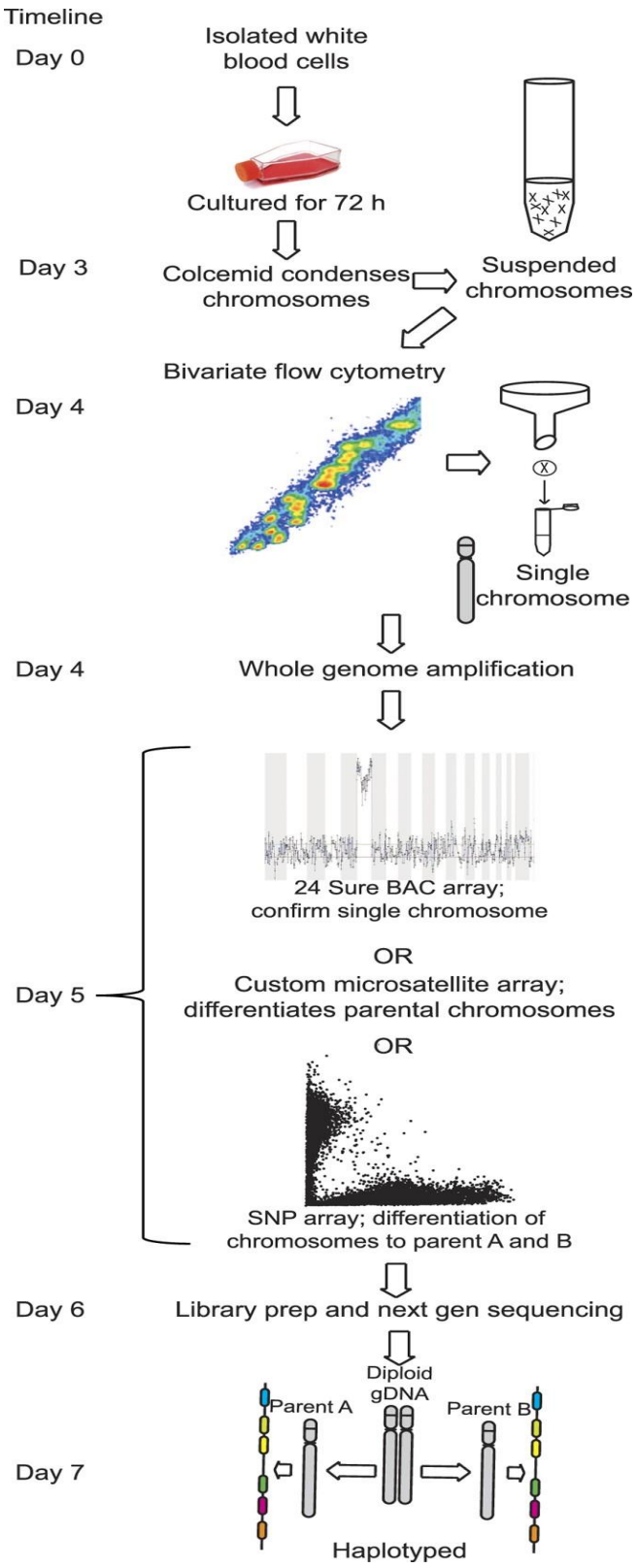
MHC genes are inherited in most cases as an **MHC haplotype**, the set of genes in a haploid genome inherited from one parent. Thus, if the parents are designated as ab and cd, then the offspring are most likely to be ac, ad, bc, or bd.

MHC- Polimorphism • MHC loci are highly polymorphic – presence of many alternative forms of the gene or allele in the population • Inherited from mother and father • New haplotypes are generated by recombination



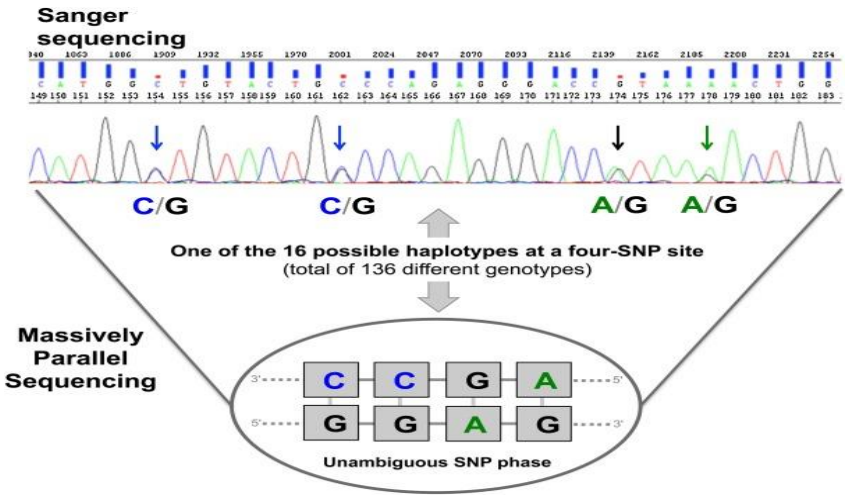
Molecular haplotyping

Genetic haplotyping and genome phasing are increasingly prominent components of both sequencing technologies and clinical genotyping/phenotyping. Chromosomal phasing or haplotype phasing refers to the physical linkage of chromosomal polymorphisms unencumbered by the ambiguity of polyploidy. The effect of phase and the quantifiable effects that the phasing of inter-chromosomal single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertions/deletions (INDELs) and structural rearrangements have on the expression of adjacent co-inherited genes is a challenging but increasingly crucial component of contemporary medicine

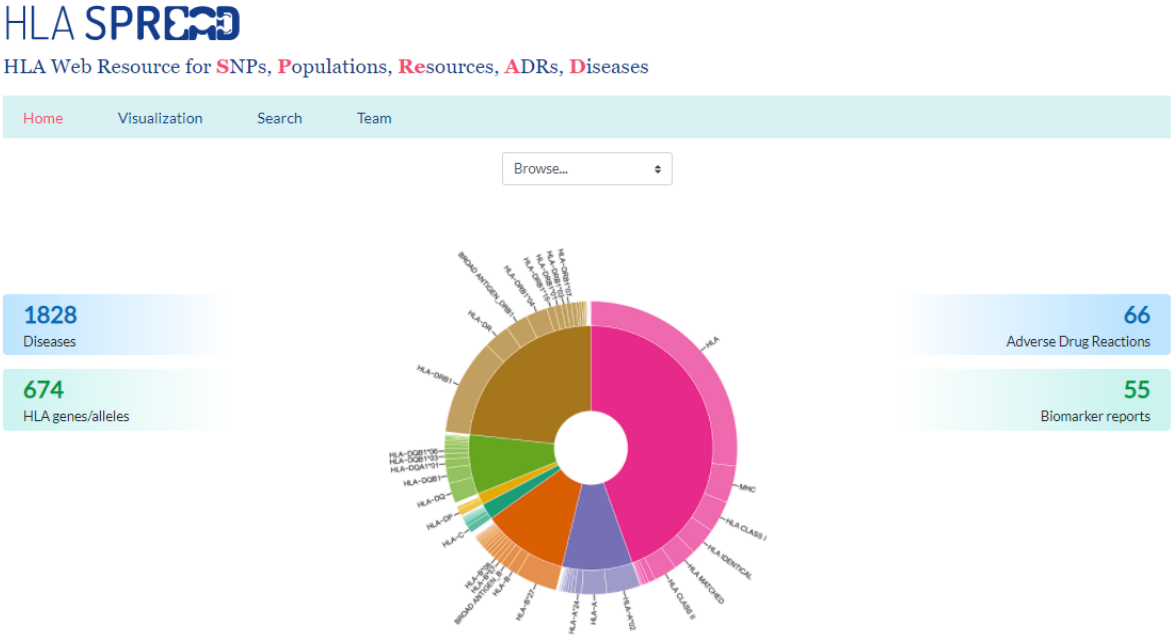

Timeline
Day 0
Isolated white blood cells
Cultured for 72 h
Day 3
Colcemid condenses chromosomes
Suspended chromosomes
Bivariate flow cytometry
Day 4
Single chromosome
Day 4
Whole genome amplification
Day 5
24 Sure BAC array; confirm single chromosome
OR
Custom microsatellite array; differentiates parental chromosomes
OR
SNP array; differentiation of chromosomes to parent A and B
Day 6
Library prep and next gen sequencing
Day 7
Diploid gDNA
Parent A
Parent B
Haplotyped

Microhaplotyping

- Microhaplotype loci (microhaps, MHs) are a novel type of molecular marker of less than 300 nucleotides, defined by two or more closely linked SNPs associated in multiple allelic combinations.
- The value of these markers is enhanced by massively parallel sequencing (MPS), which allows the sequencing of both parental haplotypes at each of the many multiplexed loci



HLA and disease associations



- The integration of computer sciences with biomedical research has accelerated the progress, both in terms of novel discoveries and data structuring.
- Natural Language Processing (NLP) is a method to extract relevant information from unstructured data.
- A simple NLP pipeline contains 4 components: data assembly, pre-processing and normalization, Named Entity Recognition (NER) and Relation Extraction (RE).
- The output of NLP algorithms, i.e. structured dataset can be used to generate insights via direct interpretation or through downstream analyses.

In recent times, NLP methods have started gaining popularity in biological sciences.

For instance, Rakhi et.al reported a text mining pipeline to study spice-disease associations and link phytochemicals from different spices/herbs to diseases.

Another report by Lee et.al highlights BioBERT, a pre-trained biomedical language representation model that can be used for various text mining tasks like Name Entity Recognition (NER), Relationship extraction (RE) and question answering, specifically on biomedical datasets.

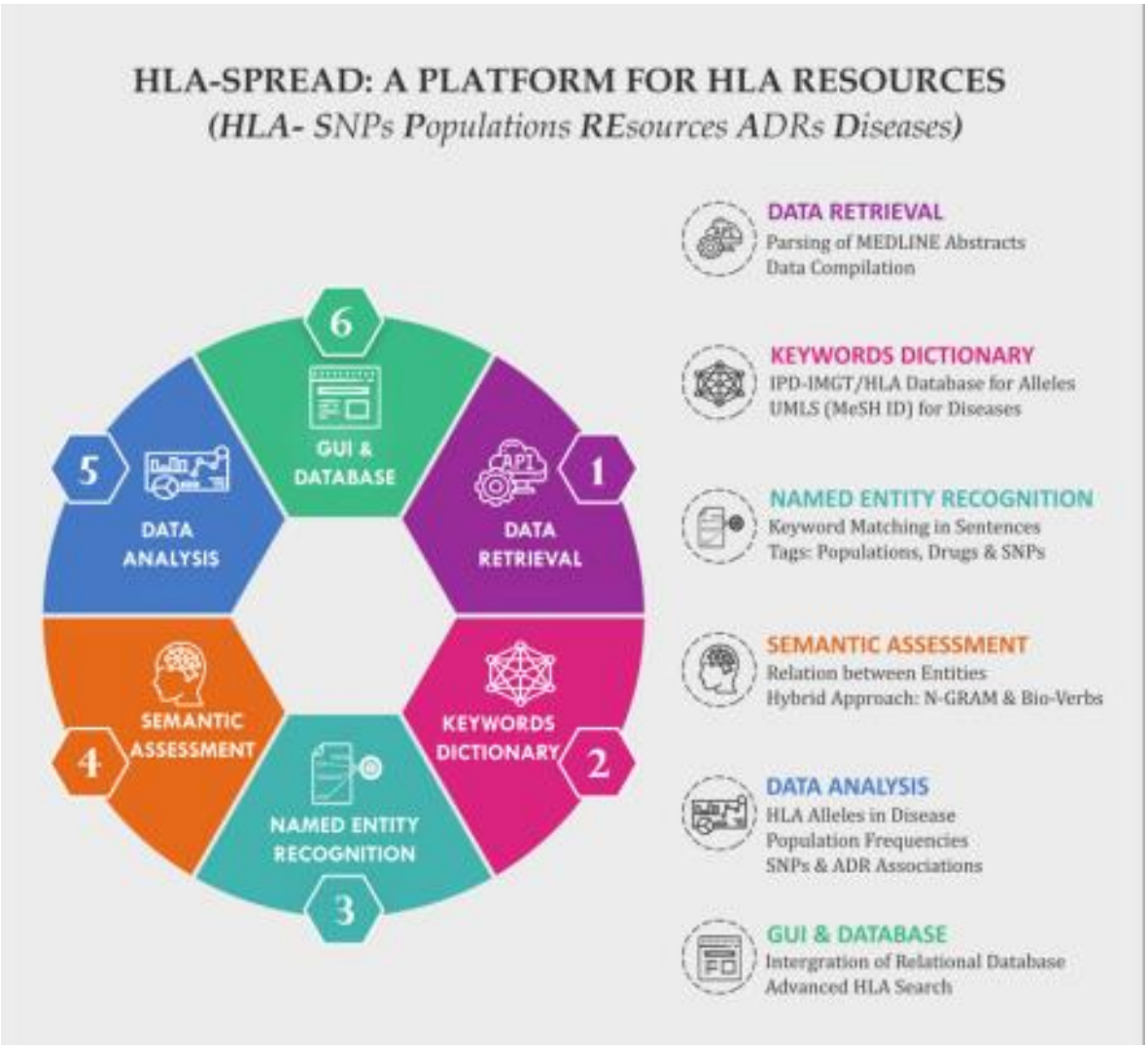
Similarly, PubTator Central is an open access tool available via NCBI that uses text mining algorithms for assisted biocuration of entities in literature. The tool uses NER to identify and thus highlight six bio-entities viz. Gene, Disease, Chemical, Mutation, Cell Line and Species from abstracts and open access articles available on PubMed.

Another interesting report by Kuleshov et.al presents a machine compiled database for studying genotype-phenotype associations generated using applications of text mining on genome-wide association studies (GWAS).

- All these resources work on similar text mining algorithms,but each has a different set of applications and tasks to perform.
- The use of these resources as such in addressing the HLA research often overlooks the extent of variability of HLA complex and involved parameters in this domain.
- For instance, PubTator Central is able to mine gene names from literature, but would not pick HLA allele information e.g. HLA-DRB1*01:01 when HLA-DRB1 is the search query.
- Conventional processes to individually mine a large amount of unstructured literature available on HLA research requires both manpower and resources.
- For understanding and integrating the observations from HLA studies we require knowledge of genomic datasets, i.e. diseases, SNPs, drugs, populations,and ethnic groups along with an understanding of the relationship between them

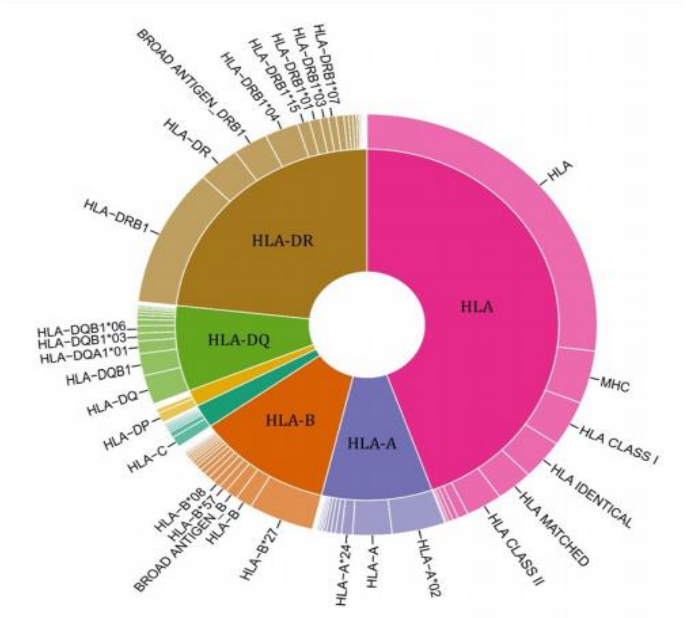
- **HLA-SPREAD** as a platform for integrated HLA resources that has been developed using NLP to understand the complexity of this locus.

- The resource provides a platform to summarize HLA related genomics knowledge as well as to design and develop new hypothesis.
- In this study, they have used publicly available ~24 million peer reviewed abstracts.
- They extracted biomedical entities including HLA alleles, diseases, SNPs, drugs and geographical locations.
- They also tried assigning positive and negative relationships between disease and alleles.
- This HLA connectivity was then used to address biologically and clinically relevant objectives like HLA-biomarkers and risk and protective alleles for various diseases.



Workflow of HLA-SPREAD:

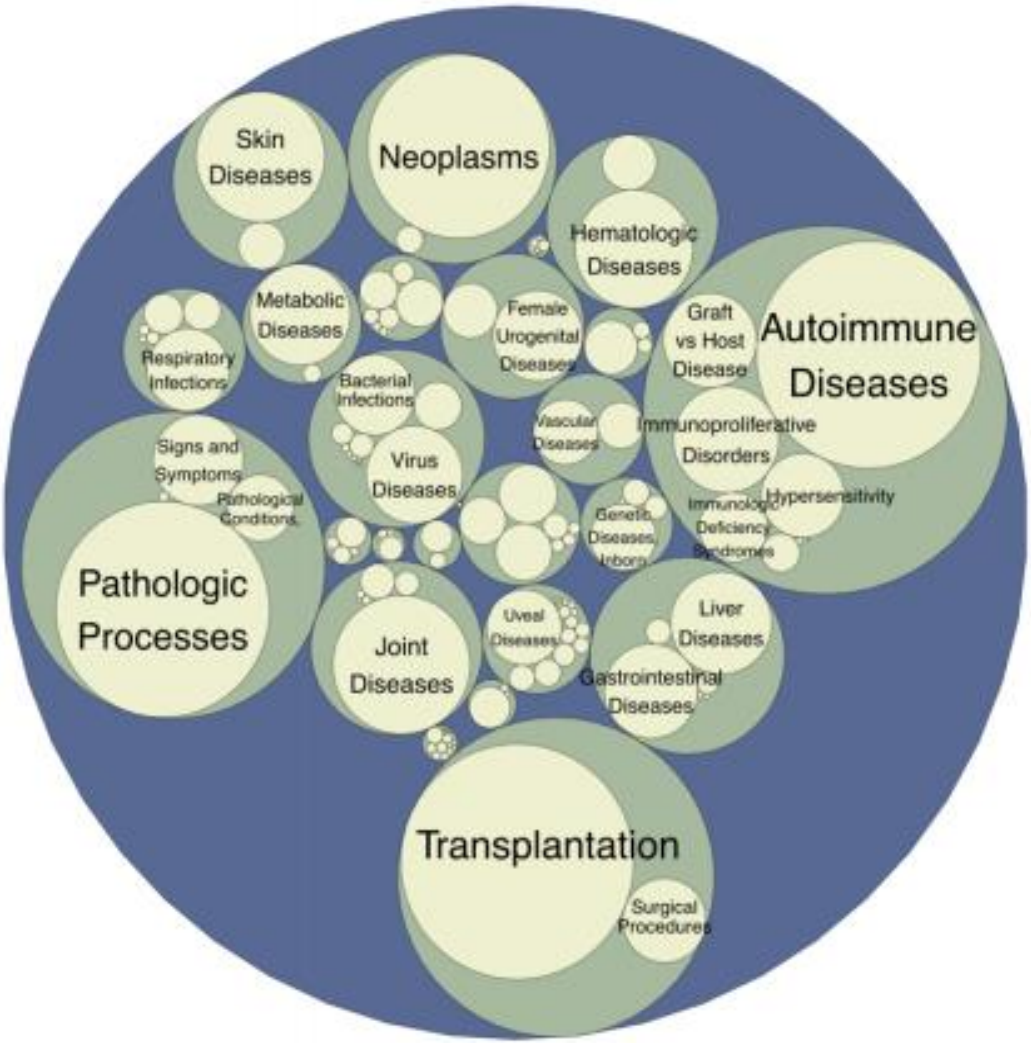
- An automated pipeline developed to extract information related from ~110,000 studies related to HLA retrieved from over 24 million abstracts.
- Structured information from these abstracts was created using Natural Language Processing methods developed into a database HLA-SPREAD.
- The **topmost reported HLA alleles associated with diseases:**
- All the HLA alleles indicated have been grouped to their second digit and represented in the pie chart.
- HLA-A, HLA-B and HLADRB1 are the most studied amongst the HLA genes
-

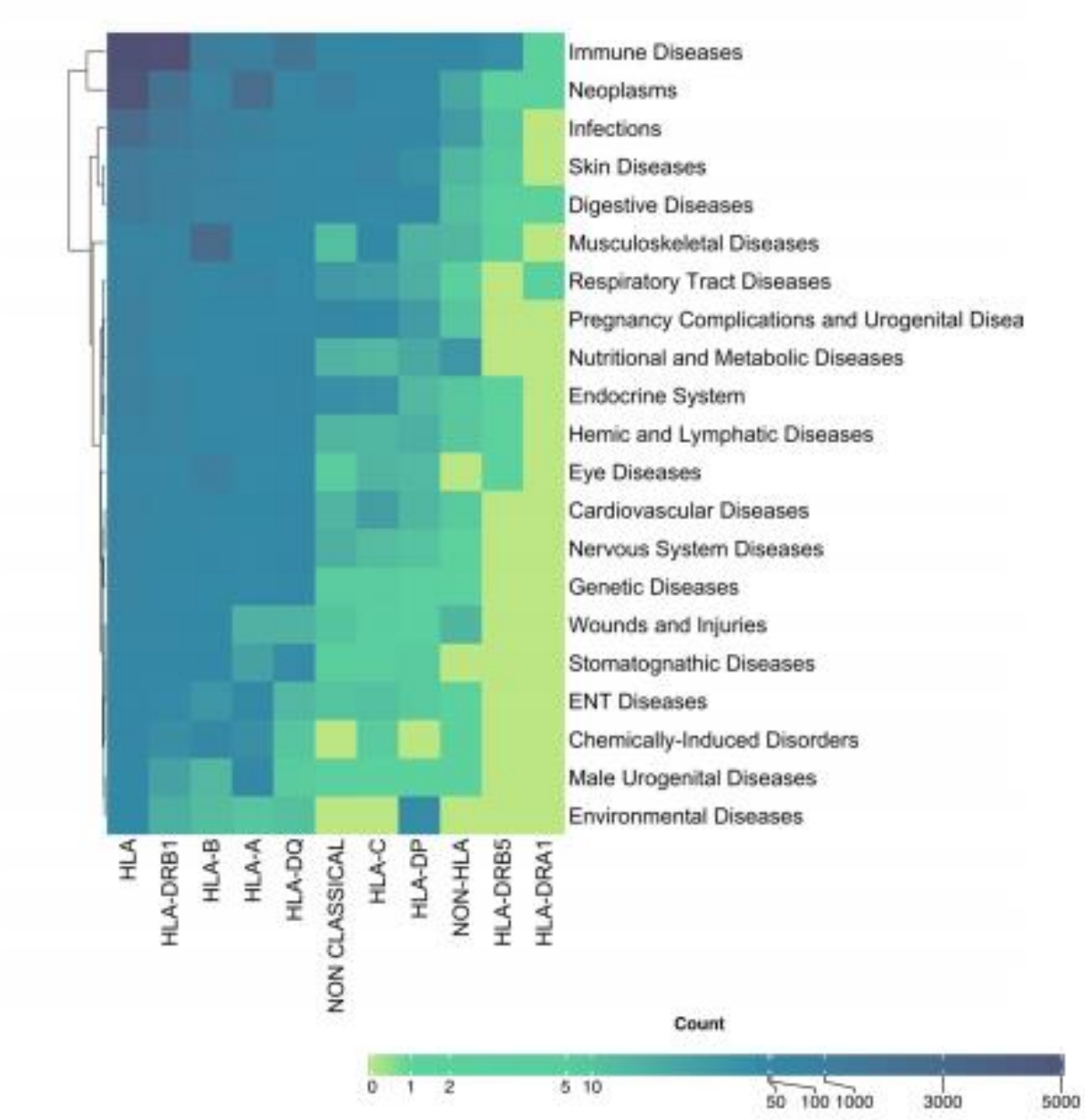


Diseases/conditions associated with HLA genes:

- Graph represents three level hierarchy of diseases.
- Each colour represents a level. There are 24 major categories as represented in green colour, which is further divided into subcategories.
- Each disease name is matched to its Mesh id and a normalised mesh keyword.
- Autoimmune, Neoplasms and Joint disease are the top most associated diseases.

- As anticipated, significant numbers of studies related to transplantation are also observed.





Heatmap of HLA Disease associations:

- The gradient heat map representing the number of diseases associated with HLA genes.
- First column represents generic “HLA” studies where specific gene information is not mentioned.
- A large number of associations were also observed with Nonclassical (HLA-E,F,G) genes

UNIT – III -IMMUNOINFORMATICS&COMPUTATIONAL VACCINOLOGY-SBIA5202

UNIT III ALLERGEN BIOINFORMATICS

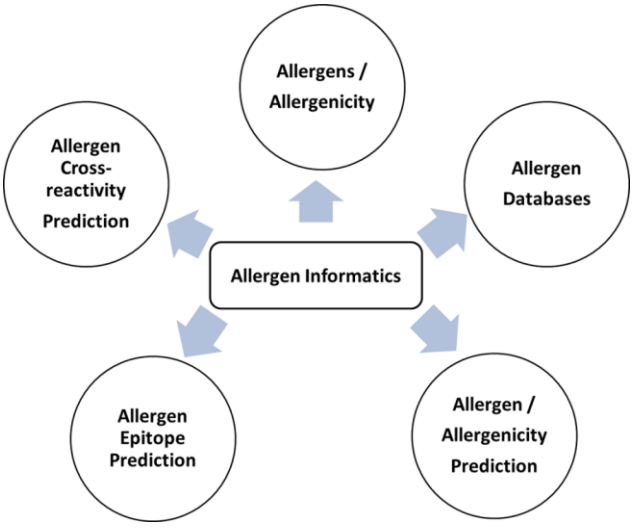
Introduction allergen database

Allergic diseases are considered as one of the major health problems worldwide due to their increasing prevalence. Advancements in genomic, proteomic, and analytical techniques have resulted in considerable progress in the field of allergology, which has led to accumulation of huge amount of data. Allergen bioinformatics comprises allergen-related data resources and computational methods/tools, which deal with an efficient archival, management, and analysis of allergological data. Significant work has been done in the area of allergen bioinformatics that has proven pivotal for the development and progress of this field. In this chapter, we describe the current status of databases and algorithms, encompassing the field of allergen bioinformatics by examining work carried out thus far with respect to features such as allergens and allergenicity, allergen databases, algorithms/tools for allergen/allergenicity prediction, allergen epitope prediction, and allergenic cross-reactivity assessment. This chapter illustrates concepts and algorithms in allergen bioinformatics, as well as it outlines the key areas for potential development in allergology field**.**

The immune system represents a very complex system comprising numerous biological molecules and processes, which combine to form body's defense against infectious agents and other threats. Immunity is basically divided into two types such as innate immunity and adaptive immunity. Innate immunity also referred to as natural, native, or nonspecific immunity acts as a first line of defense against common harmful agents. Innate immune response provides immediate protection and involves number of components such as monocytes, macrophages, neutrophils, cytokines, complement, and epithelial barriers. Adaptive or acquired immunity comprises highly specific immune responses that are elicited against particular pathogens or antigens. These immune responses are either cell mediated or antibody mediated (humoral) and executed by specialized lymphocytes or immunoglobulins, respectively. On certain occasions, the immune system produces immune responses that are harmful for the host organism. Autoimmunity denotes one such case wherein the body elicits immune responses against its own cells and tissues (self-antigens) which lead to development of autoimmune diseases. In some cases, immune system produces inappropriate immune responses known as hypersensitivity, which has deleterious effects on the

host organism. Hypersensitivity reactions are categorized into four groups based on the type of immune response and the effector mechanism involved. These are (i) immediate hypersensitivity (type I), (ii) antibody-mediated hypersensitivity (type II), (iii) immune complex–mediated hypersensitivity (type III), and (iv) cell-mediated hypersensitivity (type IV) .

Need for Allergic reactions are type I hypersensitivity reactions, which are characterized by induction of specific class of antibodies known as immunoglobulin E (IgE). These reactions are elicited against specific type of antigens commonly referred to as allergens. An allergic reaction involves specialized cells and specific molecules of the immune system. IgE antibodies induced by allergens upon allergic sensitization bind to effector cells such as basophils and mast cells via specific Fc receptors present on the surfaces of those cells. Subsequent exposure to the allergen causes cross-linking of membrane-bound IgE on these effector cells, which leads to their degranulation and release of pharmacologically active agents such as histamine. These pharmacological mediators are responsible for clinical manifestations of allergic reactions in the affected individuals. Immunogenicity in general refers to the potential of an antigen to elicit an immune response, while in case of allergens, allergenicity is considered as a reflection of its allergenic potential. Allergenicity indicates the capability of an allergen to induce clinical symptoms of allergy as well as to induce and bind to IgE antibodies. The prevalence of allergic reactions has increased significantly in the last few years, especially in the developing countries. This has resulted in considerable increase in disease burden as well as economic issues due to costs associated with these diseases. Therefore, the study of allergic diseases has gained tremendous importance as they represent one of the major health problems in urban and rural regions.



Allergens and allergenicity

Allergens represent the most critical component of an allergic reaction, although IgE antibody, Fc receptors, mast cells, and basophils as well as pharmacological mediators such as histamine and heparin also play very significant roles. Allergens are ubiquitous substances, which arise from a variety of sources such as foods, plants, animals, or environment. An allergen can either be a chemical substance (e.g., penicillin) or a protein (e.g., albumin, profilin, etc.). Majority of the allergens are proteins or glycoproteins that possess high water solubility. Several biochemical and structural features of allergens such as stability, hydrophobicity, and ligand-binding domains are known to contribute to their allergenicity . However, common molecular and structural features of allergens that are responsible for allergenicity have not yet been conclusively discovered.

Allergens are provided with a unique, unambiguous, and systematic nomenclature which has been developed and maintained by the World Health Organization (WHO) and International Union of Immunological Societies' (IUIS) "Allergen Nomenclature Sub-committee" . The nomenclature is based on the Linnean system, and an allergen, which satisfies certain biochemical and immunological criteria, is included in the WHO/IUIS nomenclature. An allergen name consists of an abbreviation of the scientific name of the allergen source organism. First 3–4 letters denote the genus name, while the subsequent 1–2 letters represent species, followed by an Arabic numeral that denotes the order of its identification. For instance, Der p 1 represents the first allergen to be characterized from the house dust mite Dermatophagoides pteronyssinus. An allergen may possess isoallergens or isoforms/variants, which are considered as multiple molecular forms of the same allergen. The WHO/IUIS nomenclature defines isoallergen as an allergen belonging to a single species, with a similar molecular size and identical biological function, and possessing ≥67% amino acid sequence identity while a variant or isoform corresponds to allergen sequences that differ by only a limited number of amino acid substitutions . It is very important to archive and study the data on isoallergens and isoforms/variants in a differentiated manner as it has been shown that variations in allergens significantly affect their allergenicity and cross-reactivity as well as influence recognition of epitopes by T cells and IgE. An allergen can be considered as a major or minor allergen based on the measure of its allergenicity. Major allergens are the ones to which

>50% of patients with an allergy to its source are sensitized, while minor allergens are recognized by a limited number of patients.

Allergens display important features such as epitopes and cross-reactivity that are very critical with respect to understanding of allergic reactions and developing newer approaches for diagnosis and treatment of allergic diseases. Epitope or antigenic determinant refers to the immunologically active region of the allergen. An epitope can be an IgE-binding epitope or a T-cell epitope depending on whether it interacts with an IgE or a T-lymphocyte. An IgE epitope can be either sequential (linear) that consists of contiguous stretch of amino acids or conformational (discontinuous), which comprises amino acids present at different loci in an antigen. An antibody is said to be cross-reactive when it recognizes and binds to multiple antigens.

IgE-binding epitopes

IgE-binding epitopes refer to the IgE recognition sites in allergens that are involved in specific interaction of allergens and IgE antibody. Inferences drawn from allergen–antibody complexes and other important studies have shown that majority of IgE-binding epitopes are conformational in nature [14]. IgE epitopes possess some defining structural and immunological features such as they are more cross-reactive in nature and have higher intrinsic flexibility. These features make them distinct from other antibody epitopes and contribute significantly in the allergenicity . Identification and in-depth analysis of IgE-binding epitopes has the potential to contribute immensely in accurate diagnosis and allergen-specific immunotherapy of allergies, especially the food allergy. Large amount of data on allergen epitopes are generated by employing strategies based on the use of overlapping synthetic peptides, recombinant allergenic fragments, cocrystal structure complexes, etc. However, it is believed that insights obtained from study of allergen–antibody complexes will be the most helpful in understanding the role these epitopes play in allergic reactions..

2.2. T-cell epitopes

T-cell epitopes are the antigenic determinants of allergens that interact with T-lymphocytes via specific T-cell receptors. T-cell epitopes of allergens have shown to be very important for the modulation of allergic response and thereby contributing to symptoms associated with allergic

diseases. They have enormous potential in the development of allergy vaccines as well as newer strategies in allergen immunotherapy, considering their fundamental role in allergic response. Recent findings have indicated that T-cell epitope repertoire in allergens is diverse than IgE epitopes, and it can be very useful in specific immunotherapy in allergy]. An analysis carried out on available epitope data has shown that T-cell epitopes are known to occur more commonly in the airborne allergens as compared to food allergens .

Cross-reactivity

Cross-reactivity denotes a clinically and immunologically critical phenomenon displayed by allergens from various sources and is the cause of pollen-food syndromes, such as the one seen in case of birch and apple. Cross-reactivity is considered as a property of antibodies and it arises when an antibody or a subgroup of antibodies recognizes more than one allergen or epitope. Two allergens are considered cross-reactive if they are recognized by a single antibody (or T-cell receptor). It has been stated that cross-reactivity among allergens at the level of B cells, T cells, and mast cells reflects clinical sensitivities and contributes very significantly in the regulation of allergic sensitization.

Cross-reactivity is predominantly an antibody defined phenomenon and IgE antibodies are shown to be more cross-reactive in nature. Affinity of the antibodies toward the allergen is known to play an important role in cross-reactivity. However, the properties of the allergenic protein are also very important and shared features on the level of both primary and tertiary structures of the cross-reactive proteins are found to be responsible for cross-reactivity . Similarity at the level of sequence is an important indicator and cross-reactivity seems to require more than 70% sequence identity. In addition to this, other factors such as the host immune response against the allergen, dosage of allergen, and mode of exposure also contribute in clinical relevance of allergic cross-reactivity. Inferences drawn from studying a large number of allergens have led to the conclusion that structural similarity among proteins from diverse sources is the molecular basis of allergic cross-reactivity. Considering the role it plays in the development of allergic symptoms, a detailed analysis of cross-reactivity has the potential to contribute in the development of new strategies in diagnosis and therapy of allergic diseases.

Allergen database

Last few years have witnessed substantial technological advances in the field of genomics and proteomics along with tremendous improvements in analytical methods. This has led to a significant progress in the area of allergy research. As a result of this, there has been a steady and continuous increase in the number of characterized protein allergens over the last few years. Efficient storage and management of data has become very important because of such incessant accumulation of molecular and clinical data on allergens. Therefore, allergen databases represent very crucial resources for basic allergy research as they are involved in archival of available allergen knowledge.

Existing allergen database:

Database	Developed by (URL)	Type of data archived	Computational tools (if any)	Updates
IUIS Allergen [36]	WHO/IUIS Allergen Nomenclature Sub-committee (http://www.allergen.org)	Sequence (isoallergens/isoforms), structure, allergenicity	–	Updated continuously
Allergome [39]	Centre for Clinical and Experimental Allergology, Italy (http://www.allergome.org)	Sequence (isoallergens/isoforms), structure, clinical, epidemiological, cross-reactivity, etc.	–	Updated continuously
Structural Database of Allergenic Proteins (SDAP) [41]	Sealy Centre for Structural Biology, University of Texas, USA (https://fermi.utmb.edu)	Sequence, structure, structural models, IgE epitopes	Yes	2013
Allergen Database For Food Safety (ADFS) [44]	National Institute of Health, Japan (http://allergen.nihs.go.jp/ADFS/)	Sequence, structure, IgE epitopes, small molecule allergens	Yes	2016
AllergenOnline [46]	Food Allergy and Resource Program (FARP) (http://www.allergenonline.org)	Sequence, allergenicity	Yes	2016

AllFam [48]	Department of Pathophysiology and Allergy Research, Medical University of Vienna, Austria (http://www.meduniwien.ac.at/allfam)	Allergen family data, cross-link to Pfam database	–	2011
AllergenPro [53]	The National Agricultural Biotechnology Information Centre, Korea (http://nabic.rda.go.kr/allergen)	Sequence, IgE epitopes	Yes	2015
AllerBase [61]	Bioinformatics Centre, Savitribai Phule Pune University, India (www.bioinfo.net.in/AllerBase/Home.html)	Sequence and structure (cross-links), IgE epitopes, IgE antibody, IgE cross-reactivity, experimental evidences of allergenicity	–	Updated continuously

Many allergen-specific databases have been developed in the past few years although they differ from each other with respect to their objectives, type of data archived, accessibility of contents, and the level of annotation and applications. In addition to dedicated allergen databases, primary bioinformatics databases also document significant data on allergens. Examples of these databases include GenBank, UniProtKB , and Protein Data Bank (PDB), which archive sequence and structure data on allergens along with its annotation.

IUIS

The IUIS Allergen Nomenclature Sub-Committee, under the auspices of the WHO, provides the systematic nomenclature of allergenic proteins and it has developed and maintained Allergen database. The database archives all of the WHO/IUIS–recognized allergens along with their isoallergens and isoforms (variants). In order to maintain a consistent allergen nomenclature for newly discovered allergens, researchers are required to submit newly described allergens to the

Allergen Nomenclature Sub-Committee before submitting their manuscript to a journal for consideration for publication.

Each allergen in this database is provided with annotation that includes biochemical name, molecular weight, information on its allergenicity, reference, etc. Additionally, sequence data for allergens and isoallergens/isoforms are also stored in the database, along with cross-references to GenBank, GenPept, and UniProtKB, as well as to PDB, for nucleotide, protein sequences, and 3D structure data, respectively. Allergen database can be searched by using allergen name, biochemical name, allergen source organism, taxonomic group, etc., as search criteria. The database is updated continuously with specific names assigned to newly discovered allergens and isoallergens/variants . Allergen database does not exemplify the comprehensive allergen data although it documents majority of the characterized allergens. This is because there are a large number of allergens that have been reported in literature which are not recognized by IUIS-Allergen. The database does not archive data on allergen epitopes and cross-reactivity.

SDAP

Structural Database of Allergenic Proteins (SDAP) is an allergen database that prominently deals with structural aspects of allergens. It houses comprehensive cross-referenced sequence data on allergens, IgE-binding epitopes, 3D structures, and models of allergens. Each allergen in SDAP is provided with cross-links to primary databases such as UniProtKB , PDB , as well as to important resources such as NCBI Taxonomy Browser and PubMed for literature references. SDAP also has a utility as a web server that integrates various computational tools, which assist structural biology–related studies dealing with allergens and their epitopes. It employs an algorithm based on the conserved properties of amino acid side chains to detect regions associated with allergenicity in novel sequences. The database consists of number of tools that can be used to assess potential cross-reactivity of allergens and also help in screening of IgE epitopes in novel proteins. SDAP does not archive complete data for allergens that are not recognized by IUIS while data on allergen cross-reactivity is also not documented.

ALLERGOME

Allergen Allergome represents an extensive repository of information on allergen molecules causing IgE-mediated (allergic, atopic) diseases. The database comprises comprehensive data on WHO/IUIS-approved allergens along with other non-recognized allergens. These allergenic molecules are selected and curated from the published literature and web-based resources. It also contains data on allergenic sources based on whether they possess identified molecules or not. Allergome documents information on allergen and isoallergens/isoforms along with their sequences. Cross-links to sequence and structure databases like UniProtKB and PDB are also provided.

Allergome can be searched by using basic and advanced search options. Basic search employs numerous search criterions such as allergen name, biochemical name, source organism, etc., while advanced search enables the user to search using specific attributes. Each allergen molecule is represented by a monograph which represents information about the three parts of allergen such as basic information, data on the native form, and its recombinant form. The most important and unique feature of Allergome platform is the presence of several support modules that deal with archival of specific aspects of allergen data. A couple of important modules are RefArray, for easy access to references stored in the Allergome, and Real Time Monitoring of IgE sensitization (ReTiME), for real-time data collection and storage of IgE sensitization data and the number of other utilities. Allergome is updated regularly and allergen data curated from literature is documented.

Allergen Database for Food Safety (ADFS) is developed as a project of the Division of Biochemistry and Immunochemistry of National Institute of Health Sciences (Japan). The aim of the database is to archive allergenic proteins and their IgE epitopes with a special emphasis on food allergens and food safety . Allergens archived in ADFS are grouped into eight categories such as pollen, mite, animal, fungus, insect, food, latex, and others, and each allergen entry is provided with the primary database accession numbers of their genes and 3D structure information. The database is also equipped with homology-based sequence search tool for the evaluation of allergenicity. One of the most distinct features of ADFS is the archival of data on small molecule, nonprotein (chemical) allergens. The database does not archive data on allergen cross-reactivity.

AllergenOnline database

AllergenOnline is a well curated allergen database that documents a peer reviewed allergen list, which is compiled from various resources such IUIS-Allergen, PubMed, scientific publications, and other allergen databases. The database was developed within the Food Allergy Research and Resource Program (FARRP) at the University of Nebraska. For each allergen, data on source organism, common name, IUIS official nomenclature, protein length, class of allergen like food allergen, contact allergen, etc., and a link to the NCBI protein (GenPept) database are provided. AllergenOnline also provides the utility for sequence-based searches for allergens, which include alignments by FASTA and an eight-amino acid short-sequence identity search. This utility can be very useful in the identification of proteins that may present a potential risk of allergenic cross-reactivity. AllergenOnline is updated every year and the last update that resulted in version 16 of the database was reported on January 27, 2016. It does not archive data on allergen epitopes as well as on allergenic cross-reactivity.

AllFam database

AllFam represents a very important resource for allergens as it is involved in classification of allergens into protein families . This study has shown that allergens are distributed into relatively few protein families and possess a limited number of biochemical functions. The structural classification of allergens in AllFam is performed by using family information from PFam and the Structural Classification of Proteins (SCOP) database , while biochemical functions of allergens were extracted from the Gene Ontology annotation database . The database provides the option of browsing lists of allergen families based on allergen source (plants, animals, and fungi) and route of exposure (inhalation, ingestion, etc.) while search for specific protein families can also be performed. Each allergen family in AllFam is linked to a family fact sheet that describes the biochemical properties of the family members as well as a list of key references related to this family. AllFam does not archive data on molecular features of individual allergens although cross-link to IUIS-Allergen and Allergome is provided for each documented allergen.

AllergenPro database

AllergenPro is a recently developed allergen database that archives data on allergen sequences, structures, and epitopes from various sources. It is an integrated database which provides information about allergens in foods, microorganisms, fungi, animals, and plants . It has been provided with a utility to search for allergens based on keywords as well as the sequence. AllergenPro is also equipped with a computational tool for the prediction of allergenicity. Prediction is based on three different approaches such as FAO/WHO guidelines (sequence)–based approach, motif-based approach, and epitope-based approach. AllergenPro does not archive data on allergen cross-reactivity while the literature references for documented allergens and epitopes have also not been provided.

Allergenicity prediction

Allergens mainly comprise commonly occurring proteins in foods, pollens, and other biological entities in the environment. It has become necessary to assess the potential allergenicity of these proteins considering the health hazards associated with allergic reactions to them. In recent years, genetic engineering and food processing methods are routinely employed for modifying the existing proteins or introducing new ones. Analysis of allergenicity of such proteins/products along with newly introduced biopharmaceuticals is absolutely essential in order to avoid transfer of an allergenic molecule. Computational assessment or prediction of allergenicity represents the major approach to test for allergenicity, and numerous bioinformatics tools/methods have been employed successfully for this purpose. The majority of these methods utilize the amino acid sequence of allergens along with its different features, while a very few approaches use structure information. Table denotes the list of computational tools/servers available for the prediction of allergens/allergenicity. In the following section, the prominent approaches used for the computational assessment or prediction of allergens/allergenicity are described briefly.

List of computational tools/servers for allergen/allergenicity prediction

No.	Method (URL)	Approach used	Efficiency
1	SDAP (http://fermi.utmb.edu/) [41]	FAO/WHO guidelines	–
2	AllergenOnline (http://www.allergenonline.org) [46]	FAO/WHO guidelines	–
3	AllergenPro (http://nabic.rda.go.kr/allergen) [53]	FAO/WHO guidelines, sequence motifs, epitopes	–
4	Allermatch (http://allermatch.org) [66]	FAO/WHO guidelines	–
5	AllerTool (http://research.i2r.a-star.edu.sg/AllerTool/) [67]	FAO/WHO guidelines, global representation of protein sequence and SVM	A_{ROC} = 0.90, SE = 86%, SP = 86%
6	WebAllergen (http://weballergen.bii.a-star.edu.sg/) [73]	Sequence motifs	–
7	AlgPred (http://www.imtech.res.in/raghava/algpred/) [75]	SVM, sequence motifs, epitopes, allergen representative peptides	Accuracy = 85%, SE = 88%, SP = 81%
8	AllerTOP (http://www.ddg-pharmfac.net/AllerTOP) [83]	Sequence based descriptors, auto and cross-covariance, machine learning	Accuracy = 85.3%, SE = 82.5%, SP = 88.1%
9	EVALLER (http://bioinformatics.bmc.uu.se/evaller.html) [86]	DFLAP algorithm and SVM	–
10	AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter/) [89]	Iterative pairwise sequence similarity and SVM	A_{ROC} = 0.928, accuracy = 95.3%, SE = 83.4%, SP = 96.4%
11	APPEL (http://jing.cz3.nus.edu.sg/cgi-bin/APPEL) [91]	Sequence based features and SVM	MCC = 0.95, SE = 93%, SP = 99.9%
12	SORTALLER (http://sortaller.gzhmu.edu.cn/) [93]	AFFP dataset, normalized BLAST *E*-values and SVM	MCC = 0.97, SE = 98.6%, SP = 98.4%
13	PREAL (http://gmobl.sjtu.edu.cn/PREAL/index.php) [96]	Biochemical and physicochemical descriptors, sequence features, subcellular locations, mRMR, SVM	Accuracy = 93.42%

13	PREAL (http://gmobl.sjtu.edu.cn/PREAL/index.php) [96]	Biochemical and physicochemical descriptors, sequence features, subcellular locations, mRMR, SVM	Accuracy = 93.42%
14	Allerdictor (http://allerdictor.vbi.vt.edu/) [99]	Sequences as text documents, Naive Bayes classifier and SVM	–
15	proAP (http://gmobl.sjtu.edu.cn/proAP/main.html) [100]	Integration of methods based on FAO/WHO guidelines, sequence motifs and SVM	–
16	AllergenFP (http://ddg-pharmfac.net/AllergenFP/) [103]	Auto and cross-covariance, descriptor-based fingerprints of residues	Accuracy = 88%, MCC = 0.759
17	FuzzyApp (http://fuzzyapp.bicpu.edu.in/fuzzyapp.php) [107]	Fuzzy rule based system	–

Sequence similarity-based approaches

One of the first studies dealing with analysis of allergenicity was put forth by Metcalfe et al. They have proposed a decision tree–based approach for allergenicity assessment of foods derived from genetically modified crops. The first computational approach for the assessment of allergenicity was provided by "Codex Alimentarius Commission" of FAO/WHO . It stated that a protein can be regarded as an allergen if it consists of an exact match with at least six contiguous amino acids or showed more than 35% similarity over a window of 80 amino acids when compared with a sequence of known allergen. This approach has been widely used to predict allergenicity and there are number of web servers for allergen prediction, which are based on it. Allermatch , AllerTool , and AllergenPro are some of the prominent web servers which employ these FAO/WHO guidelines for allergen prediction. Additionally, some of the major allergen databases such as SDAP and AllergenOnline also utilize this strategy for allergenicity prediction. A recent study performed by Verma et al. has shown that the sequence similarity-based approach gives substantially better

results when used in combination with other bioinformatics methods. However, results obtained by certain studies indicated that approaches based on these guidelines are not highly efficient for identifying allergenic proteins and many of times they lead to false or irrelevant allergenicity estimations As a result of these observations, it became necessary to discover and employ other strategies for the prediction of allergenicity.

Motif-based approaches

In a study carried out by Stadler and Stadler , it was observed that the use of sequence motifs, which represent the secondary structures of proteins, performs significantly better than the approach based on FAO/WHO guidelines. This method employs MEME motifs of a length of 50 residues for the prediction of allergenicity by using pairwise sequence alignment with certain threshold. WebAllergen is a web server for the prediction of allergenic proteins which is also based on specific detectable allergenic motifs in known allergens . Furthermore, a study carried out by Kong et al. showed that an approach based on search of multiple motifs is more specific and efficient than the conventional single motif search . AlgPred and AllergenPro are important web servers for allergen prediction in which one of the prediction approaches is based on allergen-derived motifs. A recent study that employs computational approaches for comparison of allergens and metazoan parasite proteins stated that significant sequence and structure similarity exists between parasite proteins and allergenic proteins . The analysis was carried out using sequence and structural motifs in allergens and a workflow was developed for the computational analysis of parasite proteins.

Machine learning–based approaches

Recent years have witnessed tremendous increase in the application of machine learning methods for solving biological problems. Machine learning–based approaches have been widely used for predicting various aspects of protein function . These methods are also employed routinely for the development of algorithms to predict allergenicity of novel proteins.

Although Support Vector Machine (SVM) is the most commonly used machine learning method for allergen prediction, other methods have also been frequently employed. One of the earliest methods was developed by Zorzet et al. that utilizes a k-Nearest-Neighbor (kNN) classification algorithm for the prediction of allergenicity, while a Bayesian classifier was employed by Soeria-Atmadja et

al. for the same purpose. An approach based on the combination of hidden Markov model (HMM) and conserved motifs in allergen was also used to successfully predict protein allergenicity . Dimitrov et al. developed two artificial neural network (ANN)-based algorithms for allergenicity prediction, which utilize descriptors derived from amino acids that denote their structural and physicochemical properties. AllerTOP is an online bioinformatics tool to perform the computational prediction of allergens . This algorithm employs descriptors that denote the chemical properties of amino acids in allergen sequences and auto- and cross-covariance transformation along with five machine learning methods for classification. These methods are random forest, multilayer perceptron, logistic regression, decision tree, naïve Bayes, and kNN.

There are number of web-based tools/servers developed which use SVM for performing classification/prediction of allergens. AlgPred is one of the earlier web servers developed for the prediction of allergenic proteins . It employs SVM with amino acid and dipeptide composition as features of allergens to achieve accuracy of 85.02 and 84.00%, respectively. EVALLER is another web server created for in silico determination of potential allergenicity with very good efficiency . It performs detection based on filtered length–adjusted allergen peptides (DFLAP) algorithm and SVM. AllerTool web server also applies SVM-based algorithm for the prediction of allergenicity and provides sensitivity and specificity of 86.00% . AllerHunter is an important web-based computational system for allergenicity assessment which uses a scheme based on iterative pairwise sequence similarity encoding along with SVM. The method is very efficient with a sensitivity of 83.4% and a specificity of 96.4%.

A web-based tool APPEL is developed for the prediction of allergenic proteins that employs physicochemical and structural features derived from allergen sequence in combination with SVM. Zhang et al. have developed an online allergen prediction tool titled SORTALLER, which is based on allergen family featured peptide (AFFP) dataset and employs SVM as a classifier. An algorithm developed by Mohabatkar et al. for the prediction of allergenic proteins utilizes pseudoamino acid composition (PseAAC) along with SVM and provides an accuracy of 91.19%. PREAL is web-based tool that performs allergen prediction by using SVM along with feature selection methods such as maximum relevance minimum redundancy (mRMR) and incremental feature selection (IFS). A combination of hydrophobicity amino acid index and discrete Fourier transform along with an SVM classifier is employed for highly efficient prediction of allergenicity in a signal-processing bioinformatics approach . Allerdictor is web server that specializes in large-scale allergen

discovery. It models protein sequences as text documents and employs SVM in text classification for carrying out allergen prediction.

Other approaches

A study carried out by Wang et al. evaluated sequence-, motif-, and SVM-based approaches for the computational prediction of allergens and also performed parameter optimization to obtain better performance . The resulting methods from this study are integrated and made available as a web application titled proAP . AllergenFP is a recently developed web server for allergenicity prediction that utilizes alignment-free descriptor-based fingerprint approach . The descriptors used here are important properties of amino acid such as size, hydrophobicity, relative abundance, helix, and beta-strand forming propensities, etc. In a structure-based approach proposed by Bragin et al. information derived from protein 3D structure is used for the representation of protein surface as patches designated as discontinuous peptides. It is observed that prediction of allergenic proteins based on this approach gave better accuracy. Vijayakumar and Lakshmi have developed a fuzzy inference system–based algorithm for allergenicity prediction that utilizes five different modules . These modules consist of a machine learning classifier, motif search, sequence similarity, FAO/WHO evaluation scheme, etc. FuzzyApp , a web server based on fuzzy rule–based system, is then developed for the prediction of allergenicity . Jiang et al. performed an analysis of food allergens using a computational model that simulates gastric fluid digestion. This study stated that food allergens could be classified as alimentary canal-sensitized and nonalimentary canal-sensitized allergens based on the digestibility of these allergens in simulated gastric fluid.

Computational prediction of allergen epitopes

Epitopes represent distinctive amino acid residues on the antigens and are important determinants of an immune response. Identification of epitopes is considered a key aspect of designing highly effective multiple-subunit vaccines and developing efficient diagnostic and therapy methods against allergens. Although experimental methods have been very useful for the identification of epitopes, their usefulness is restricted because of their time- and cost-intensive nature and inability in dealing with large-scale elucidation of epitopes. Hence, computational approaches are considered to be very beneficial alternative as they are cost and time effective.

Large number of highly efficient algorithms and tools have been developed over the years for the computational prediction of epitopes. These methods deal with the prediction of both B-cell and T-cell epitopes as well as sequential (linear) and discontinuous (conformational) epitopes. Based on the information (data) utilized for performing prediction, the methodologies can be grouped as sequence-based or structure-based approaches. Many sequence-based linear epitope prediction methods for B cells have been developed and used since long time and majority of them are propensity scale and machine learning–based methods. Some of the major tools/servers that deal with the prediction of linear B-cell epitopes are listed in Table

No.	Method (URL)	Approach used	Efficiency
1	ABCPred (http://www.imtech.res.in/raghava/abcpred/) [111]	Fixed length epitope patterns, recurrent ANNs	Accuracy = 65.93%, SE = 67.14%, SP = 64.71%
2	APCpred (http://ccb.bmi.ac.cn/APCpred/) [112]	Amino acid anchoring pair composition (APC) and SVM	A_{ROC} = 0.809, accuracy = 72.94%
3	BCPreds (http://ailab.cs.iastate.edu/bcpreds/) [113]	SVM classifiers with string kernels	A_{ROC} = 0.758
4	BcePred (http://www.imtech.res.in/raghava/bcepred/) [114]	Physicochemical properties of epitope residues	Accuracy = 58.7%
5	BepiPred (http://www.cbs.dtu.dk/services/BepiPred/) [115]	Parker's hydrophilicity scale and HMM	–
6	BEST (http://biomine.ece.ualberta.ca/BEST/) [116]	Antigen sequence features, SVM	A_{ROC} = 0.85

6	BEST (http://biomine.ece.ualberta.ca/BEST/) [116]	Antigen sequence features, SVM	A_{ROC} = 0.85
7	Bayesb (http://www.immunopred.org/bayesb/index.html) [117]	Bayes feature extraction and SVM	A_{ROC} = 0.84, accuracy = 74.5%
8	COBEpro (http://scratch.proteomics.ics.uci.edu) [118]	Antigen fragment score and SVM	A_{ROC} = 0.829
9	EPMLR (http://www.bioinfo.tsinghua.edu.cn/epitope/EPMLR/) [119]	Sequence features and multiple linear regression (MLR)	A_{ROC} = 0.728, SE = 81.8%, SP = 64.1%
10	IEDB Analysis Resource (http://tools.iedb.org/bcell/) [120]	A collection of tools based on various methods	–
11	LBtope (http://www.imtech.res.in/raghava/lbtope/) [121]	Large datasets of epitopes, KNN, SVM	Accuracy = 86%
12	SVMTriP (http://sysbio.unl.edu/SVMTriP/) [122]	Tri-peptide similarity, propensity scores and SVM	A_{ROC} = 0.702, SE = 80.1%, SP = 55.2%

Number of methods that utilize 3D structure of antigens for discontinuous epitope prediction have also been developed. These methods use different approaches for prediction such as solvent accessibility of surface residues, solvent accessibility with propensity scores, and propensity scores with packing density of amino acids. An account of major tools/servers that are involved in conformational epitope prediction is provided in Table

No.	Method (URL)	Approach used	Efficiency
1	BEpro (formerly PEPITO) (http://pepito.proteomics.ics.uci.edu/) [127]	3D structure of antigen, amino acid propensity scores	A_{ROC} = 0.75
2	B-Pred (http://immuno.bio.uniroma2.it/bpred) [128]	3D structure or model of antigen, solvent exposure of residues	SE = 0.70
3	CBTOPE (http://www.imtech.res.in/raghava/cbtope/) [129]	Sequence features and SVM	A_{ROC} = 0.9, Accuracy = 85%
4	CEP (http://196.1.114.49/cgi-bin/cep.pl) [124]	3D structure of antigen, solvent accessibility of amino acids	Accuracy = 75%
5	DiscoTope 2.0 (http://www.cbs.dtu.dk/services/DiscoTope/) [125]	3D structure of antigen, epitope propensity scores, surface accessibility	A_{ROC} = 0.824
6	ElliPro (http://tools.immuneepitope.org/tools/ElliPro) [130]	3D structure of antigen, Thornton's method, residue clustering algorithm	A_{ROC} = 0.732
7	Epitopia (http://epitopia.tau.ac.il/) [131]	Antigen sequence or 3D structure, Naïve Bayes classifier	A_{ROC} = 0.59
8	EPSVR (http://sysbio.unl.edu/EPSVR/) [132]	3D structure of antigen, Support vector regression (SVR)	A_{ROC} = 0.597
9	EPMeta (http://sysbio.unl.edu/EPMeta/) [132]	Meta server integrating EPSVR with other methods	A_{ROC} = 0.638
10	EPCES (http://sysbio.unl.edu/EPCES/) [133]	3D structure of antigen, surface features	–
11	SEPPA 2.0 (http://badd.tongji.edu.cn/seppa/) [126]	3D structure of antigen, subcellular localization of antigen, residue propensity, etc.	A_{ROC} = 0.745

Studies have shown that the analysis of antigen–antibody complex structures is very useful for the characterization of conformational epitopes. A dedicated resource titled AgAbDb that archives the interactions derived from antigen–antibody complexes is available, which can be very useful for the analysis of epitopes . Several algorithms have also been developed for the prediction of T-cell epitopes in antigens. These methodologies deal with the prediction of peptides that possess the ability to interact with specific major histocompatibility complex (MHC) molecules. Machine

learning–based approaches are very commonly employed for this purpose and are found to be very efficient. The details of epitope prediction methods/tools for B cells and T cells have been reviewed elsewhere . Some of the important tools/servers that perform the prediction of T-cell epitopes are listed in Table 5. Recently, it has been shown that epitope prediction can be performed over the whole proteome by integrating multiple epitope prediction methods. Antibody-specific epitope prediction has emerged as a significant alternative to the traditional antibody-independent epitope prediction methods

No.	Method (URL)	Approach used	Efficiency
1	CTLPred (http://www.imtech.res.in/raghava/ctlpred/index.html) [140]	Cytotoxic T-lymphocyte epitopes, SVM, ANN	Accuracy = 75.2%
2	EpiJen (http://www.ddg-pharmfac.net/epijen/EpiJen/EpiJen.htm) [141]	Multi-step algorithm that employs integrated approach	Accuracy = 60%
3	EpiTOP (http://www.pharmfac.net/EpiTOP/) [142]	QSAR approach based on proteochemometrics	Accuracy = 89%
4	NetCTLpan (http://www.cbs.dtu.dk/services/NetCTLpan/) [143]	Integrated method employing proteasomal cleavage, TAP transport efficiency, and MHC class I binding affinity	A_{ROC} = 0.95
5	NetMHCIIpan-3.0 (http://www.cbs.dtu.dk/services/NetMHCIIpan-3.0/) [144]	A method for all HLA class II molecules based on peptide-binding MHC environment	A_{ROC} = 0.807
6	PREDIVAC (http://predivac.biosci.uq.edu.au/) [145]	Based on specificity-determining residues	–
7	SYFPEITHI (http://www.syfpeithi.de/bin/MHCServer.dll/EpitopePrediction.htm) [146]	Scoring system based on position of residue in the epitopes	–
8	TEPITOPEpan (http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/) [147]	Algorithm based on HLA-DR binding pocket similarity	–
9	WAPP (http://abi.inf.uni-tuebingen.de/Services/WAPP/) [148]	Combination of methods based on proteasomal cleavage, TAP transport and MHC binding	–

Computational prediction of allergenic cross-reactivity

Cross-reactivity plays an important role in allergic reaction from the immunological and clinical context. Therefore, the computational prediction of allergenic cross-reactivity has been considered of substantial significance. The prediction of cross-reactivity in allergens is associated with the prediction of allergenicity for the majority of the cases. This is mainly because the antigenic determinants that contribute to the cross-reactivity in allergens are also responsible for their allergenicity. As a result of this, many of the tools/algorithms that have been developed for the prediction of allergens/allergenicity also perform cross-reactivity prediction.

The criteria defined by FAO/WHO experts, which have been mentioned earlier, help to identify cross-reactivity in allergens . AllerTool is a web server that performs cross-reactivity prediction based on amino acid sequence and WHO/FAO guidelines . It also provides a graphical representation of the published and predicted cross-reactivity patterns of allergens. Stadler and Stadler developed a sequence-based approach and stated that motif-based strategy provides better results for the computational assessment of cross-reactivity than the FAO/WHO guidelines. SDAP , which is a specialized allergen database described before, also comprises a sequence-based tool for the identification of cross-reactivity among allergens. AllerHunter is a SVM-based web server that deals with efficient assessment of allergic cross-reactivity in proteins. A recently developed fuzzy inference system–based algorithm for allergenicity prediction is also able to predict cross-reactivity in allergens

UNIT – IV-IMMUNOINFORMATICS&COMPUTATIONAL VACCINOLOGY-SBIA5202

UNIT IV COMPUTATIONAL VACCINOLOGY

Vaccines –types

A **vaccine** is made from very small amounts of weak or dead germs that can cause diseases — for example, viruses, bacteria, or toxins. It prepares your body to fight the disease faster and more effectively so you won't get sick.

Example: Children younger than age 13 need 2 doses of the chickenpox **vaccine**.

Vaccination

Vaccination is the act of getting a vaccine, usually as a shot.

Immunization

Immunization is the process of becoming immune to (protected against) a disease.

Example: Because of continued and widespread **immunization** in the United States, it's rare for Americans to get polio.

Immunization can also mean the process of getting vaccinated. For example, your "immunization schedule," is the timing of your shots.

Vaccine Types

There are several different types of vaccines. Each type is designed to teach your immune system how to fight off certain kinds of germs—and the serious diseases they cause.

When scientists create vaccines, they consider:

- How your immune system responds to the germ
- Who needs to be vaccinated against the germ
- The best technology or approach to create the vaccine

Based on a number of these factors, scientists decide which type of vaccine they will make. There are several types of vaccines, including:

- Inactivated vaccines
- Live-attenuated vaccines
- Messenger RNA (mRNA) vaccines
- Subunit, recombinant, polysaccharide, and conjugate vaccines
- Toxoid vaccines

- Viral vector vaccines

Inactivated vaccines

Inactivated vaccines use the killed version of the germ that causes a disease.

Inactivated vaccines usually don't provide immunity (protection) that's as strong as live vaccines. So you may need several doses over time (booster shots) in order to get ongoing immunity against diseases.

Inactivated vaccines are used to protect against:

- Hepatitis A
- Flu (shot only)
- Polio (shot only)
- Rabies

Live-attenuated vaccines

Live vaccines use a weakened (or attenuated) form of the germ that causes a disease.

Because these vaccines are so similar to the natural infection that they help prevent, they create a strong and long-lasting immune response. Just 1 or 2 doses of most live vaccines can give you a lifetime of protection against a germ and the disease it causes.

But live vaccines also have some limitations. For example:

Because they contain a small amount of the weakened live virus, some people should talk to their health care provider before receiving them, such as people with weakened immune systems, long-term health problems, or people who've had an organ transplant.

They need to be kept cool, so they don't travel well. That means they can't be used in countries with limited access to refrigerators.

Live vaccines are used to protect against:

- Measles, mumps, rubella (MMR combined vaccine)

- Rotavirus
- SmallpoxExternal Link: You are leaving vaccines.gov and entering a non-federal website. View full disclaimer.
- Chickenpox
- Yellow fever

Messenger RNA vaccines—also called mRNA vaccines

Researchers have been studying and working with mRNA vaccinesExternal Link: You are leaving vaccines.gov and entering a non-federal website. View full disclaimer. for decades and this technology was used to make some of the COVID-19 vaccines. mRNA vaccines make proteins in order to trigger an immune response. mRNA vaccines have several benefits compared to other types of vaccines, including shorter manufacturing times and, because they do not contain a live virus, no risk of causing disease in the person getting vaccinated.

mRNA vaccines are used to protect against:

COVID-19

Subunit, recombinant, polysaccharide, and conjugate vaccines

Subunit, recombinant, polysaccharide, and conjugate vaccines use specific pieces of the germ—like its protein, sugar, or capsid (a casing around the germ).

Because these vaccines use only specific pieces of the germ, they give a very strong immune response that's targeted to key parts of the germ. They can also be used on almost everyone who needs them, including people with weakened immune systems and long-term health problems.

One limitation of these vaccines is that you may need booster shots to get ongoing protection against diseases.

These vaccines are used to protect against:

- Hib (Haemophilus influenzae type b) disease
- Hepatitis B
- HPV (Human papillomavirus)
- Whooping cough (part of the DTaP combined vaccine)

- Pneumococcal disease
- Meningococcal disease
- Shingles

Toxoid vaccines

Toxoid vaccines use a toxin (harmful product) made by the germ that causes a disease. They create immunity to the parts of the germ that cause a disease instead of the germ itself. That means the immune response is targeted to the toxin instead of the whole germ.

Like some other types of vaccines, you may need booster shots to get ongoing protection against diseases.

Toxoid vaccines are used to protect against:

- Diphtheria
- Tetanus

Viral vector vaccines

For decades, scientists studied viral vector vaccinesExternal Link: You are leaving vaccines.gov and entering a non-federal website. View full disclaimer.. Some vaccines recently used for Ebola outbreaks have used viral vector technology, and a number of studies have focused on viral vector vaccines against other infectious diseases such as Zika, flu, and HIV. Scientists used this technology to make COVID-19 vaccines as well.

Viral vector vaccines use a modified version of a different virus as a vector to deliver protection. Several different viruses have been used as vectors, including influenza, vesicular stomatitis virus (VSV), measles virus, and adenovirus, which causes the common cold. Adenovirus is one of the viral vectors used in some COVID-19 vaccines being studied in clinical trials. Viral vector vaccines are used to protect against:

COVID-19

There are various types of vaccines that are routinely given to children

Attenuated (weakened) live viruses- These vaccines contain a live virus that has been weakened during the manufacturing process so that they do not cause the actual disease in the person being vaccinated. However, because they contain a small amount of the weakened live virus, people with

weakened immune systems should talk to their healthcare provider before receiving them. Examples include vaccines that prevent chickenpox and rotavirus and measles, mumps and rubella.

Inactivated (killed) viruses- These vaccines contain a virus that has been killed so as not to cause disease, but the body still recognizes it and stimulates production of antibodies against the virus. They can be given to individuals with weakened immune systems. Examples include vaccines to prevent polio and hepatitis A.

Subunits- In some cases, the entire virus or bacteria is not required for an immune response to prevent disease; just the important parts, a portion or a "subunit" of the disease-causing bacteria or virus is needed to provide protection. The vaccine to prevent influenza (the flu) that is given as a shot is an example of a subunit vaccine, because it is made with parts of the influenza virus.

Toxoids- Some bacteria cause illness in people by secreting a poison (a toxin). Scientists discovered that weakening the toxins, so that they are “detoxified” does not cause illness. Examples of vaccines that contain toxoids include those to prevent tetanus and diphtheria disease.

Recombinant- These vaccines are made by genetic engineering, the process and method of manipulating the genetic material of an organism. An example of this type of vaccine is those that prevent certain diseases caused by the human papillomavirus (HPV), such as cervical cancer. In this case, the genes that code for a specific protein from each of the virus types of HPV included in the vaccine are expressed in yeast to create large quantities of the protein. The protein that is produced is purified and then used to make the vaccine. Because the vaccine only contains a protein, and not the entire virus, the vaccine cannot cause the HPV infection. It is the body's immune response to the recombinant protein(s) that then protects against diseases caused by the naturally occurring virus.

Polysaccharides- To protect against certain disease-causing bacteria, the main antigens in vaccine are sugar-like substances called polysaccharides; these are purified from the bacteria to make polysaccharide vaccines. However, vaccines composed solely of purified polysaccharides are only effective in older children and adults. Pneumovax 23, a vaccine for the prevention of pneumococcal disease caused by 23 different strains, is an example of a polysaccharide vaccine.

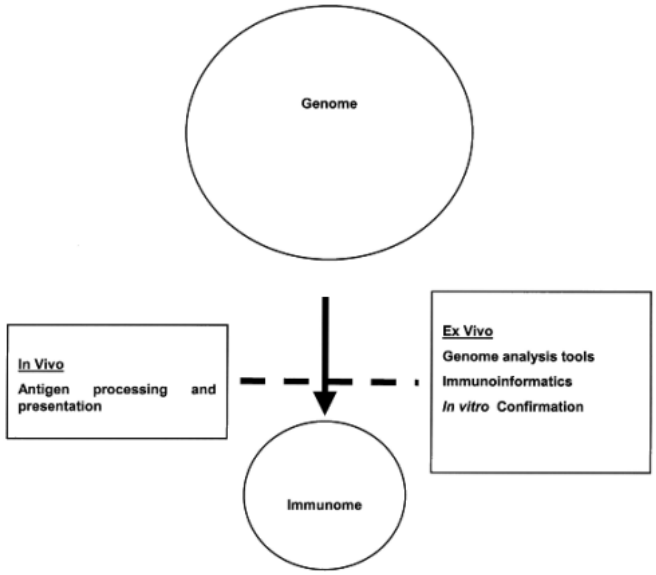
Conjugates- Vaccines made only with polysaccharides do not work very well in young children because their immune system has not fully developed. To make vaccines that protect young children against diseases caused by certain bacteria, the polysaccharides are connected to a protein so that the immune system can recognize and respond to the polysaccharide. The protein acts as a "carrier" for the part of the vaccine that will make protective antibodies in the body. Examples of conjugate vaccines include those to prevent invasive disease caused by Haemophilus influenzae type b (Hib).

From immunome to vaccine- epitope mapping

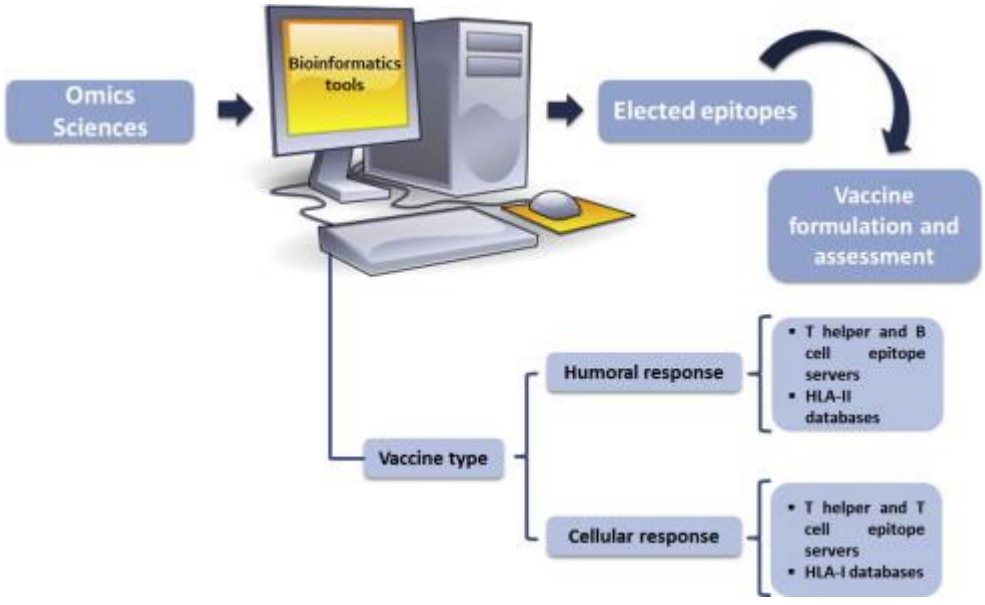
De¢ning the immunome

In general, host immune response to a pathogen is thought to be due to a number of pathogen-speci¢c responses (provided by antibodies; T helper cells, which drive antibody response; and CTL, for intracellular pathogens). The T cell response is stimulated by the presence of short peptides or epitopes, that are derived from pathogen-speci¢c antigens by antigen presenting cells and presented to T cells in the context of MHC surface proteins (major histocompatability complex molecules, or MHC). Whether the immune response is directed against a single immunodominant epitope or against many epitopes, the generation of a protective immune response does not appear to require the development of T and B cell memory to every possible peptide from every antigen in the entire pathogen. T and B cell responses to the ensemble of epitopes derived from selected antigens (and not to the whole pathogen) appear su⁄cient to provide protective immunity. Consider for example the hepatitis B virus (HBV) vaccine, the cowpox virus, known as vaccinia, which is used prevent smallpox infection, and BCG vaccine, which is used to prevent TB disease. The HBV vaccine consists of a single recombinant protein, separated from the other proteins of HBV. Antibodies developed in response to this protein-based vaccine protect against hepatitis B infection. Thus only HBV protein, and not the entire virus, is needed to generate a protective immune response. While the smallpox and TB and their vaccines are related, they are not identical. Presumably the protective immune response against the pathogen that is generated by immunization with vaccinia is due to B and T cell epitopes that are conserved between the pathogen and its vaccine. Therefore, vaccines that contain a single protein (HBV vaccine) or a subset of proteins (vaccinia and BCG), or even just epitopes derived from those proteins, may be able to create an immune response to

challenge the pathogen that is just as e¡ective as vaccines containing whole proteins or whole pathogens. The set of epitopes, which de¢ne the ‘immunome’ of the pathogen, can be de¢ned and discovered by comparing genome sequences and applying new immunoinformatics tools



Approach	*Result*
Compare unrelated genomes	Uncover similarities that could be used to make broader vaccines
Contrast genomes of virulent and avirulent strains	Identify antigens associated with virulence for vaccines
Compare genomes of microbial pathogens and their vaccines, or vaccines that might be used	Identify antigens responsible for protective immune response (e.g. BCG and Mtb)

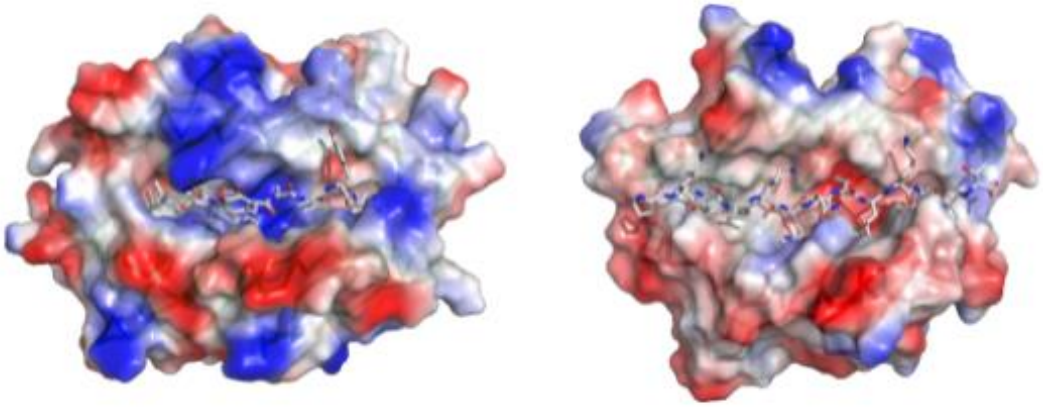


T-Cell Epitope Prediction

T-cell epitope prediction aims to identify the shortest peptides within an antigen that are able to stimulate either CD4 or CD8 T-cells [7]. This capacity to stimulate T-cells is called immunogenicity, and it is confirmed in assays requiring synthetic peptides derived from antigens [5, 6]. There are many distinct peptides within antigens and T-cell prediction methods aim to identify those that are immunogenic. T-cell epitope immunogenicity is contingent on three basic steps: (i) antigen processing, (ii) peptide binding to MHC molecules, and (iii) recognition by a cognate TCR. Of these three events, MHC-peptide binding is the most selective one at determining T-cell epitopes

Prediction of Peptide-MHC Binding

MHC I and MHC II molecules have similar 3D-structures with bound peptides sitting in a groove delineated by two α-helices overlying a floor comprised of eight antiparallel β-strands. However, there are also key differences between MHC I and II binding grooves that we must highlight for they condition peptide-binding predictions (Figure 3). The peptide-binding cleft of MHC I molecules is closed as it is made by a single α chain. As a result, MHC I molecules can only bind short peptides ranging from 9 to 11 amino acids, whose N- and C-terminal ends remain pinned to conserved residues of the MHC I molecule through a network of hydrogen bonds [10, 11]. The MHC I peptide-binding groove also contains deep binding pockets with tight physicochemical preferences that facilitate binding predictions. There is a complication however. Peptides that have different sizes and bind to the same MHC I molecule often use alternative binding pockets [12]. Therefore, methods predicting peptide-MHC I binding require a fixed peptide length. However, since most MHC I peptide ligands have 9 residues, it is generally preferable to predict peptides with that size. In contrast, the peptide-binding groove of MHC II molecules is open, allowing the N- and C-terminal ends of a peptide to extend beyond the binding groove [10, 11]. As a result, MHC II-bound peptides vary widely in length (9–22 residues), although only a core of nine residues (peptide-binding core) sits into the MHC II binding groove. Therefore, peptide-MHC II binding prediction methods often target to identify these peptide-binding cores. MHC II molecule binding pockets are also shallower and less demanding than those of MHC I molecules. As a consequence, peptide-binding prediction to MHC II molecules is less accurate than that of MHC I molecules.



MHC molecule binding groove. The figure depicts the molecular surface as seen by the TCR of representative MHC I and II molecules. Note how the binding groove of the MHC I molecule is closed but that of MHC II is open. As a result, MHC I molecules bind short peptides (8–11 amino acids), while MHC II molecules bind longer peptides (9–22 amino acids).

Method used for prediction of peptide-MHC binding. Keys for methods: SM: sequence motif; SB: structure-based; MM: motif matrix; QAM: quantitative affinity matrix; SVM: support vector machine; ANN: artificial neural network; QSAR: quantitative structure-activity relationship model; combined: tool uses different methods including ANN and QAM, selecting the more appropriate method for each distinct MHC molecule. The table also indicates whether the tools predict quantitative binding affinity (A), supertypes (S), TAP binding (T), and proteasomal cleavage (P); marked with an X in the affirmative case.

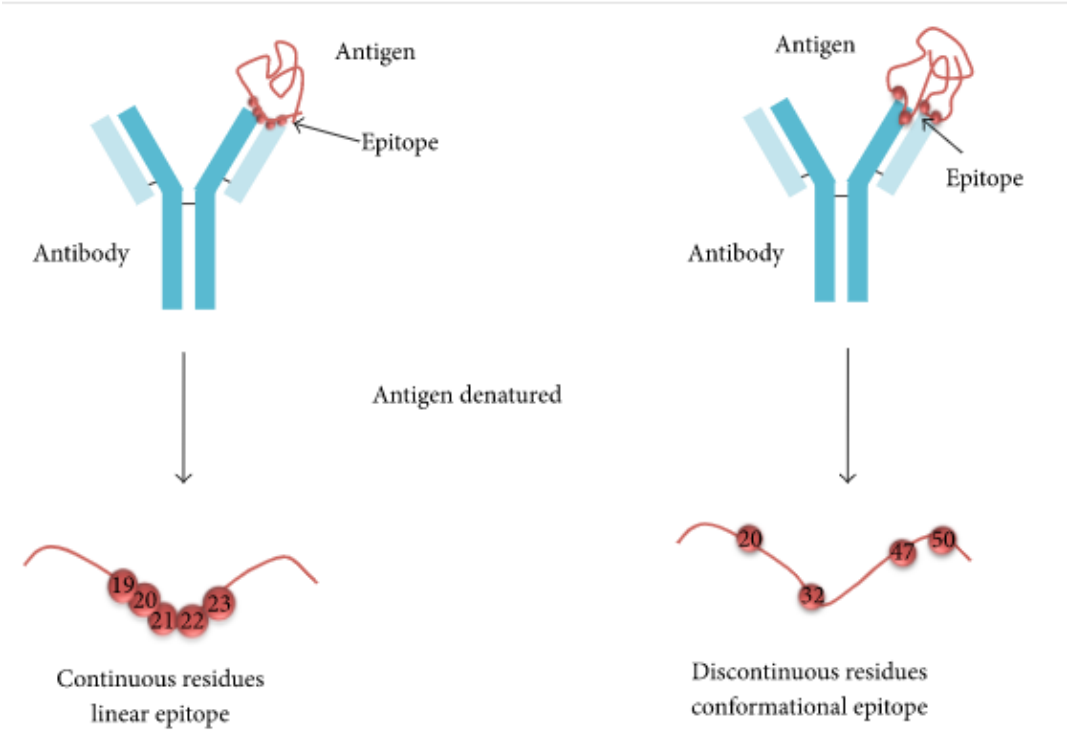
Tool	Method[1]	MHC class	A	S	T	P
EpiDOCK	SB	II	—	—	—	—
MotifScan	SM	I and II	—	X	—	—
Rankpep	MM	I and II	—	—	—	X
SYFPEITHI	MM	I and II	—	—	—	—
MAPPP	MM	I	—	X	—	X

PREDIVAC	MM	II	—	—	—	—
PEPVAC	MM	I	—	X	—	X
EPISOPT	MM	I	—	X	—	—
Vaxign	MM	I and II	—	—	—	—
MHCPred	QSAR	I and II	X	—	—	—
EpiTOP	QSAR	II	X	—	—	—
BIMAS	QAM	I	X			
TEPITOPE	QAM	II	X	—	—	—
Propred	QAM	II	X	X	—	—
Propred-1	QAM	I	X	X	—	X
EpiJen	QAM	I	X	—	X	X
IEDB-MHCI	Combined	I	X	—	—	—
IEDB-MHCII	Combined	II	X	—	—	—
IL4pred	SVM	II	—	—	—	—
MULTIPRED2	ANN	I and II	—	X	—	—
MHC2PRED	SVM	II	—	—	—	—
NetMHC	ANN	I	X	—	—	—
NetMHCII	ANN	II	X	—	—	—
NetMHCpan	ANN	I	X	—	—	—

NetMHCIIpan	ANN	II	X	—	—	—
nHLApred	ANN	I	—	—	—	X
SVMHC	SVM	I and II	—	—	—	—
SVRMHC	SVM	I and II	X	—	—	—
NetCTL	ANN	I	X	X	X	X
WAPP	SVM	I	—	—	X	X

. Prediction of B-Cell Epitopes

B-cell epitope prediction aims to facilitate B-cell epitope identification with the practical purpose of replacing the antigen for antibody production or for carrying structure-function studies. Any solvent-exposed region in the antigen can be subject of recognition by antibodies. Nonetheless, B-cell epitopes can be divided in two main groups: linear and conformational (Figure). Linear B-cell epitopes consist of sequential residues, peptides, whereas conformational B-cell epitopes cosist of patches of solvent-exposed atoms from residues that are not necessarily sequential (Figure). Therefore, linear and conformational B-cell epitopes are also known as continuous and discontinuous B-cell epitopes, respectively. Antibodies recognizing linear B-cell epitopes can recognize denatured antigens, while denaturing the antigen results in loss of recognition for conformational B-cell epitopes. Most B-cell epitopes (approximately a 90%) are conformational and, in fact, only a minority of native antigens contains linear B-cell epitopes. We will review both, prediction of linear and conformational B-cell epitopes



Linear and conformational B-cell epitopes. Linear B-cell epitopes (a) are composed of sequential/continuous residues, while conformational B-cell epitopes (b) contain scattered/discontinuous residues along the sequence.

Prediction of Linear B-Cell Epitopes

Linear B-cell epitopes consist of peptides which can readily be used to replace antigens for immunizations and antibody production. Therefore, despite being a minority, prediction of linear B-cell epitopes have received major attention. Linear B-cell epitopes are predicted from the primary sequence of antigens using sequence-based methods. Early computational methods for the prediction of B-cell epitopes were based on simple amino acid propensity scales depicting physicochemical features of B-cellepitopes. For example, Hopp and Wood applied residue hydrophilicity calculations for B-cell epitope prediction [96, 97] on the assumption that hydrophilic regions are predominantly located on the protein surface and are potentially antigenic. We know now, however, that protein surfaces contain roughly the same number of hydrophilic and hydrophobic residues [98]. Other amino acid propensity scales introduced for B-cell epitope prediction are based on flexibility [99], surface accessibility [100], and β-turn propensity [101]. Current available bioinformatics tools to predict linear B-cell epitopes using propensity scales include PREDITOP [102] and PEOPLE [103] (Table 2). PREDITOP [102] uses a multiparametric

algorithm based on hydrophilicity, accessibility, flexibility, and secondary structure properties of the amino acids. PEOPLE [103] uses the same parameters and in addition includes the assessment of β-turns. A related method to predict B-cell epitopes was introduced by Kolaskar and Tongaonkar [104], consisting on a simple antigenicity scale derived from physicochemical properties and frequencies of amino acids in experimentally determined B-cell epitopes. This index is perhaps the most popular antigenic scale for B-cell epitope prediction, and it is actually implemented by GCG [105] and EMBOSS [106] packages. Comparative evaluations of propensity scales carried out in a dataset of 85 linear B-cell epitopes showed that most propensity scales predicted between 50 and 70% of B-cell epitopes, with the β-turn scale reaching the best values [101, 107]. It has also been shown that combining the different scales does not appear to improve predictions [102, 108]. Moreover, Blythe and Flower 09] demonstrated that single-scale amino acid propensity scales are not reliable to predict epitope location.

Tool	Method
Linear B cell epitope	
PEOPLE	Propensity scale method
BepiPred	ML (DT)
ABCpred	ML (ANN)
LBtope	ML (ANN)
BCPREDS	ML (SVM)
SVMtrip	ML (SVM)
Conformational B-cell epitope	
CEP	Structure-based method (solvent accessibility)
DiscoTope	Structure-based method (surface accessibility and propensity amino acid score)

ElliPro	Structure-based method (geometrical properties)
PEPITO	Structure-based method (physicochemical properties and geometrical structure)
SEPPA	Structure-based method (physicochemical properties and geometrical structure)
EPITOPIA	Structure-based method (ML-naïve Bayes)
EPSVR	Structure-based method (ML-SVR)
EPIPRED	Structure-based method (ASEP, Docking)
PEASE	Structure-based method (ASEP, ML)
MIMOX	Mimotope
PEPITOPE	Mimotope
EpiSearch	Mimotope
MIMOPRO	Mimotope
CBTOPE	Sequence based (SVM)

Tools in immunoinformatics for the prediction binding affinity between peptide:

MHC:Peptide Binding Prediction – SYFPEITHI

SYFPEITHI is a database comprising more than 7000 peptide sequences known to bind class I and class II MHC molecules. The entries are compiled from published reports only.

This Database contains information on:

- Peptide sequences
- anchor positions
- MHC specificity
- source proteins, source organisms

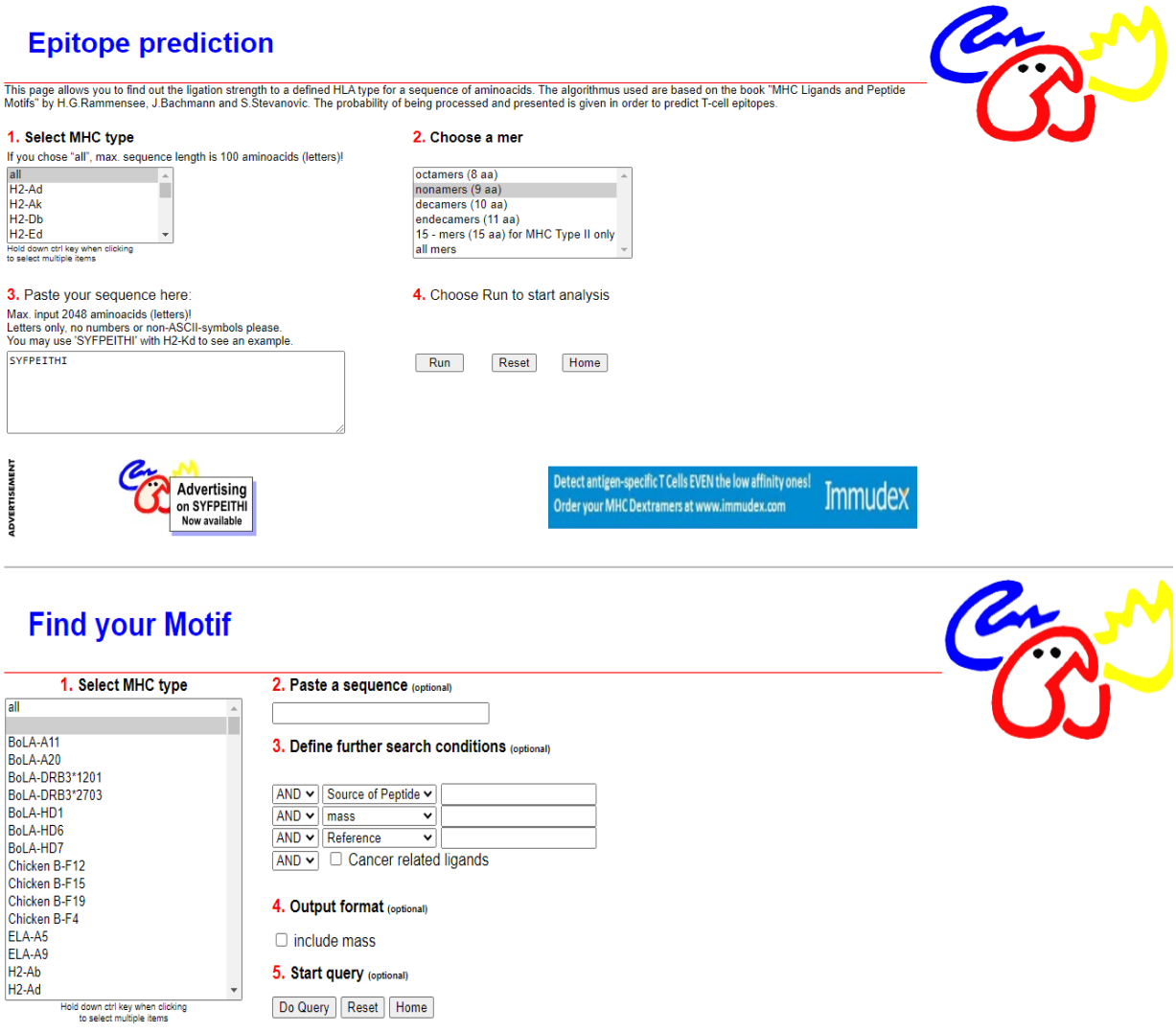
publication references

Links with sequence databases and 'MedLine' are available online

Epitope prediction and retrieval of sequences according to their molecular mass is also possible

SYFPEITHI and its scores

The scoring system evaluates every amino acid within a given peptide. Individual amino acids may be given the arbitrary value 1 for amino acids that are only slightly preferred in the respective position, optimal anchor residues are given the value 15; any value between these two is possible. Negative values are also possible for amino acids which are disadvantageous for the peptide's binding capacity at a certain sequence position. The allocation of values is based on the frequency of the respective amino acid in natural ligands, T-cell epitopes, or binding peptides



All data are stored centrally in a relational client-server database system (RDBMS). The main table of the RDBMS contains examples of ligands and T-cell epitopes, as well as additional information on the specific role (anchor and auxiliary anchor amino acids) and position of each individual amino acid (aa). When dataare browsed, the information is transformed to present a formatted version in which anchor aa are given in bold letters and auxiliary anchors are underlined. The main table is linked in a manyto-one relationship to the list of sources, and in a many-to-many relationship with the table of references. Each record in the table of source proteins refers to the specific EMBL ID, whereas each entry in the table of references is linked to the accession number (AN) of the reference in the NLM-PubMed database.Database retrieval can be performed on any HTML-browser supporting JavaScript. The main page of the database (http:/

/www.uni-tuebingen.de/uni/kxi/) offers three sections: “Find Your Motif”, “Epitope prediction” and “Information”. After a preselection of one or multiple MHC-types, the “Find Your Motif” section allows the user to search for a complete or truncated sequence of up to nine aa, a given peptide source, or a reference.

The search can be narrowed down even further by choosing a specific aa on a given position as anchor or auxiliary anchor. All search criteria may also be combined to obtain a complex analysis. When a search is performed, an SQL-query is generated and the results are presented on a dynamically composed HTML page. The page of results lists the MHC type, motifs, peptide sources, and references. From each peptide source and each reference, a hyperlink to the EMBL or PubMed database is generated, respectively.

The algorithm used for epitope prediction is written in Object Pascal. In brief, a two-dimensional data array is built up, where the letters of the aa represent the row index and the pocket numbers represent the column index. The scores in the array-cells of the matrix shown in Table 1 can be addressed directly by a pair of indices. Starting at the first aa, the sequence is then divided into octa-, nona- or decamers and for each oligomer the sum of the scores of the aa contained is calculated. The process is then repeated until the end of the sequence is reached. Amino acids that frequently occur in anchor positions are given the value 10, the value 8 is given to amino acids present in a significant number of ligands, and 6 for rarely occurring residues; amino acids of auxiliary anchor positions are given the value 6, less frequent residues of the same set have a

coefficient of 4; preferred amino acids have coefficients of 1–4 according to the strength of signals in pool sequencing or the occurrence in individual sequences. Amino acids that are regarded as unfavorable for binding have a coefficient of –1 to –3. These values are taken into account in the algorithm.

MHCPred

MHCPred uses the **additive method** to predict the binding affinity of major histocompatibility complex (**MHC**) class I and II molecules and also to the Transporter associated with Processing (TAP). Allele specific Quantitative Structure Activity Relationship (**QSAR**) models were generated using partial least squares (PLS).

How to use the query form:

Sequence:

Sequences are limited to a maximum of 1000 residues. Longer sequences are not be accepted, for reasons of CPU usage, so please check before submitting. Currently only sequences in plain format are accepted: other formats, such as FASTA or GCG, are not accepted by MHCPred.

Select the allele:

Class I MHC	Class II MHC	Other
HLA_A*0101	DRB1*0101	TAP
HLA_A*0201	DRB1*0401	
HLA_A*0202	DRB1*0701	
HLA_A*0203		
HLA_A*0206		
HLA_A*1101		
HLA_A*0301		
HLA_A*3301		
HLA_A*6801		
HLA_A*6802		

HLA_B*3501

H-2Db

H-2Kb

H-2Kk

Select the model:

Two models are used for binding affinity prediction. The first uses only amino acids contribution; the second is based on the contributions of amino acids and their interactions**.**

The cut-off value of IC50:

Suggested IC50 values are between 0.01 to 5000 nM. If the value is above 5000, then the peptide is unlikely to bind MHC molecules.

Select the anchor positions:

The user can select specific residues at specific anchor positions. The maximum number of selected anchor positions is 4.

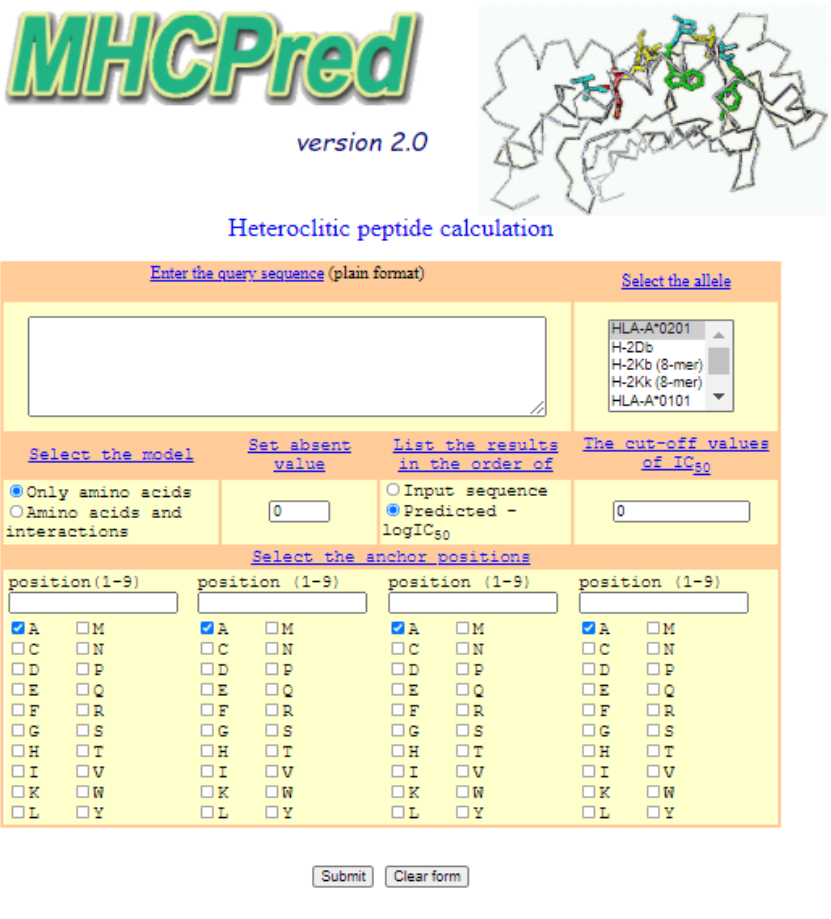
Note: The positions have to be different each time, i.e. the user cannot enter the same position twice.

Results:

The results of the program is shown in a table with three columns. The first column shows the peptide sequences, the second and the third column show the predicted IC50 and IC50 values respectively. If the IC50 value is above 5000 then the peptide will not bind to the MHC molecules. The order of the peptides is sorted by the IC50 values. Peptides with lower IC50 values (or higher predicted IC50 value) are listed first, and the non-binders are listed at the bottom of the table. To view the position of an interested peptide, the user can click on the peptide and the query sequence will be shown in a separate window, with the peptide in question highlighted.

Background:

nM = 10-9 M. Molar is traditionally used by chemists to describe the concentration of chemical solutions. "molar" describes a concentration in moles per liter (mol/L). A solution described as 1.0 µM has a concentration of 1.0 µmol/L. Mole is the SI fundamental unit of the amount of a substance (as distinct from its mass or weight). Moles measure the actual number of atoms or molecules in an object. The official definition, adopted as part of the SI system in 1971, is that one mole of a substance contains just as many elementary entities (atoms, molecules, ions, or other kinds of particles) as there are atoms in 12 grams of carbon 12. The actual number of "elementary entities" in a mole is called Avogadro's number after the Italian chemist and physicist Amedeo Avogadro (1776-1856). Careful measurement determines Avogadro's number to be approximately 602.214 199 x 1021 entities per mole.



Servers for peptide–MHC binding

Server name	Class	URL
SYFPEITHI	I and II	http://syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/EpiPredict.htm
BIMAS	I	http://bimas.dcrt.nih.gov/molbio/hla_bind/
MHC-THREAD	II	http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/
EpiPredict	II	http://www.epipredict.de/index.html
HLA-DR4 binding	II	http://www-dcs.nci.nih.gov/branches/surgery/sbprog.html
ProPred	II	http://www.imtech.res.in/raghava/propred/
RankPep	I and II	http://www.mifoundation.org/Tools/rankpep.html
SVMHC	I	http://www.sbc.su.se/svmhc/
PREDEP	I	http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/
NetMHC	I	http://www.cbs.dtu.dk/services/NetMHC/
PREDICT	I	http://sdmc.krdl.org.sg:8080/predict/
LpPep	I	http://reiner.bu.edu/zhiping/lppep.html

TEPITOPE

Accurate identification of peptides binding to specific Major Histocompatibility Complex Class II (MHC-II) molecules is of great importance for elucidating the underlying mechanism of immune recognition, as well as for developing effective epitope-based vaccines and promising immunotherapies for many severe diseases. Due to extreme polymorphism of MHC-II alleles and the high cost of biochemical experiments, the development of computational methods for accurate prediction of binding peptides of MHC-II molecules, particularly for the ones with few or no experimental data, has become a topic of increasing interest. TEPITOPE is a well-used computational approach because of its good interpretability and relatively high performance. However, TEPITOPE can be applied to only 51 out of over 700 known HLA DR molecules.

TEPITOPE has a library of 11 PSSMs. One PSSM is a 20×9 matrix where nine binding specificity vectors correspond to nine pockets. Each of the 11 PSSMs corresponds to one of 11 known DRB alleles. TEPITOPEpan uses this library to generate a PSSM for an arbitrary HLA-DRB allele. In a

generated PSSM, each vector is a weighted average of binding specificity vectors of the corresponding pocket over 11 DRB alleles. The weight can be computed by pocket sequence similarity. Thus the assumption behind TEPITOPEpan is that different alleles have similar binding preferences for one pocket (e.g. P1) if their MHC amino acids for the pocket are similar. The procedure of TEPITOPEpan has the following three steps:

Step 1: Generating pseudosequences of MHC binding pockets.

Step 2: Computing the pocket similarity and weight between alleles.

Step 3: Computing PSSM.

NetMHC

Motivation: Many biological processes are guided by receptor interactions with linear ligands of variable length. One such receptor is the MHC class I molecule. The length preferences vary depending on the MHC allele, but are generally limited to peptides of length 8 to 11 amino acids. On this relatively simple system, we developed a sequence alignment method based on artificial neural networks that allows insertions and deletions in the alignment. **Results:** We show that prediction methods based on alignments that include insertions and deletions have significantly higher performance than methods trained on peptides of single lengths. Also, we illustrate how the location of deletions can aid the interpretation of the modes of binding of the peptide-MHC, as in the case of long peptides bulging out of the MHC groove or protruding at either terminus. Finally, we demonstrate that the method can learn the length profile of different MHC molecules, and quantified the reduction of the experimental effort required to identify potential epitopes using our prediction algorithm. **Availability:** The NetMHC-4.0 method for the prediction of peptide-MHC class I binding affinity using gapped sequence alignment is publicly available at: http://www.cbs.dtu.dk/services/NetMHC-4.0.

1. Specify the input sequences

All the input sequences must be in one-letter amino acid code. The alphabet is as follows (case sensitive):

A C D E F G H I K L M N P Q R S T V W Y and X (unknown)

Any other symbol will be converted to X before processing.

The server allows for input in either FASTA or PEPTIDE format.

Sequences can be submitted in the following two formats:

Paste a single sequence (just the amino acids) or a number of sequences in FASTA format or a list of peptides into the upper window of the main server page.

Select a FASTA or PEPTIDE file on your local disk, either by typing the file name into the lower window or by browsing the disk.

At most 5000 sequences per submission; each sequence not more than 20,000 amino acids and not less than 8 amino acids.

2. Customize your run

1. Specify peptide length (only for FASTA input). **By default input proteins are digested into 9-**mer peptides.

2. Select species/loci from the scroll-down menu.

3. Select allele(s) from the scroll-down menu or type in the allele names separated by commas (without blank spaces). If you choose to type in the allele names, you can consult the List of MHC molecule names.; use the molecule names in the first column.

4. Optionally specify thresholds for strong and weak binders. They are expressed in terms of %Rank, that is percentile of the predicted binding affinity compared to the distribution of affinities calculated on set of 400.000 random natural peptides. The peptide will be identified as a strong binder if it is found among the top x% predicted peptides, where x% is the specified threshold for strong binders (by default 0.5%). The peptide will be identified as a weak binder if the % Rank is above the threshold of the strong binders but below the specified threshold for the weak binders (by default 2%).

5. Tick the box Sort by affinity to have the output sorted by descending predicted binding affinity.

3. Submit the job

Click on the "Submit" button. The status of your job (either 'queued' or 'running') will be displayed and constantly updated until it terminates and the server output appears in the browser window.

At any time during the wait you may enter your e-mail address and simply leave the window. Your job will continue; when it terminates you will be notified by e-mail with a URL to your results. They will be stored on the server for 24 hours

4. Output

A description of the output format can be found HERE.

Output format

DESCRIPTION

The prediction output for each molecule consists of the following columns:

Pos Residue number (starting from 0)

HLA Molecule/allele name

Peptide Amino acid sequence of the potential ligand

Core The minimal 9 amino acid binding core directly in contact with the MHC

Offset The starting position of the Core within the Peptide (if > 0, the method predicts a N-terminal protrusion)

I_pos Position of the insertion, if any.

I_len Length of the insertion.

D_pos Position of the deletion, if any.

D_len Length of the deletion.

iCore Interaction core. This is the sequence of the binding core including eventual insertions of deletions.

Identity Protein identifier, i.e. the name of the Fasta entry.

1-log50k(aff) Log-transformed binding affinity. Some reference transformations: 50,000nM -> logAff=0; 500nM -> logAff=0.426; 50nM -> logAff=0.638; 1nM -> logAff=1.000.

Affinity(nM) Predicted binding affinity in nanoMolar units.

%Rank Rank of the predicted affinity compared to a set of 400.000 random natural peptides. This measure is not affected by inherent bias of certain molecules towards higher or lower mean predicted affinities. Strong binders are defined as having %rank<0.5, and weak binders with %rank<2. We advise to select candidate binders based on %Rank rather than nM Affinity

BindLevel (SB: strong binder, WB: weak binder). The peptide will be identified as a strong binder if the % Rank is below the specified threshold for the strong binders, by default 0.5%. The peptide will be identified as a weak binder if the % Rank is above the threshold of the strong binders but below the specified threshold for the weak binders, by default 2%.

EXAMPLE OUTPUT

Fasta input:

>Gag_180_209

TPQDLNTMLNTVGGHQAAMQMLKETINEEA

Peptide length: 8 and 9

Allele: HLA-A*0301

will return the following predictions:

NetMHC version 4.0

Input is in FSA format

Peptide length 8,9

Affinity Threshold for Strong binding peptides 50.000

Affinity Threshold for Weak binding peptides 500.000

Rank Threshold for Strong binding peptides 0.500

Rank Threshold for Weak binding peptides 2.000

---

pos	HLA	peptide	Core	Offset	I_pos	I_len	D_pos	D_len	iCore	Identity	1-log50k(aff)	Affinity(nM)
0	HLA-A0301	TPQDLNTM	-TPQDLNTM	0	0	1	0	0	TPQDLNTM	Gag_180_209	0.014	43017.00
1	HLA-A0301	PQDLNTML	PQDLNTML-	0	8	1	0	0	PQDLNTML	Gag_180_209	0.021	39881.02
2	HLA-A0301	QDLNTMLN	-QDLNTMLN	0	0	1	0	0	QDLNTMLN	Gag_180_209	0.018	41073.47
3	HLA-A0301	DLNTMLNT	DLN-TMLNT	0	3	1	0	0	DLNTMLNT	Gag_180_209	0.019	40552.86
4	HLA-A0301	LNTMLNTV	-LNTMLNTV	0	0	1	0	0	LNTMLNTV	Gag_180_209	0.035	34098.43
5	HLA-A0301	NTMLNTVG	NTMLNTVG-	0	8	1	0	0	NTMLNTVG	Gag_180_209	0.025	38038.41
6	HLA-A0301	TMLNTVGG	TMLNTVGG-	0	8	1	0	0	TMLNTVGG	Gag_180_209	0.034	34544.05
7	HLA-A0301	MLNTVGGH	MLNTV-GGH	0	5	1	0	0	MLNTVGGH	Gag_180_209	0.083	20462.88
8	HLA-A0301	LNTVGGHQ	-LNTVGGHQ	0	0	1	0	0	LNTVGGHQ	Gag_180_209	0.018	41270.38
9	HLA-A0301	NTVGGHQA	NTVGGHQA-	0	8	1	0	0	NTVGGHQA	Gag_180_209	0.015	42434.54
10	HLA-A0301	TVGGHQAA	TVGGHQAA-	0	8	1	0	0	TVGGHQAA	Gag_180_209	0.021	39642.67
11	HLA-A0301	VGGHQAAM	-VGGHQAAM	0	0	1	0	0	VGGHQAAM	Gag_180_209	0.021	39730.28
12	HLA-A0301	GGHQAAMQ	GGHQAAMQ-	0	8	1	0	0	GGHQAAMQ	Gag_180_209	0.015	42652.28
13	HLA-A0301	GHQAAMQM	G-HQAAMQM	0	1	1	0	0	GHQAAMQM	Gag_180_209	0.020	40135.11
14	HLA-A0301	HQAAMQML	HQAAMQML-	0	8	1	0	0	HQAAMQML	Gag_180_209	0.057	27116.52
15	HLA-A0301	QAAMQMLK	-QAAMQMLK	0	0	1	0	0	QAAMQMLK	Gag_180_209	0.238	3800.57
16	HLA-A0301	AAMQMLKE	AAM-QMLKE	0	3	1	0	0	AAMQMLKE	Gag_180_209	0.021	39659.42
17	HLA-A0301	AMQMLKET	AMQMLKET-	0	8	1	0	0	AMQMLKET	Gag_180_209	0.019	40509.00
18	HLA-A0301	MQMLKETI	MQMLKET-I	0	7	1	0	0	MQMLKETI	Gag_180_209	0.033	35088.76
19	HLA-A0301	QMLKETIN	QMLKETIN-	0	8	1	0	0	QMLKETIN	Gag_180_209	0.029	36469.85
20	HLA-A0301	MLKETINE	MLKETINE-	0	8	1	0	0	MLKETINE	Gag_180_209	0.027	37444.68
21	HLA-A0301	LKETINEE	-LKETINEE	0	0	1	0	0	LKETINEE	Gag_180_209	0.011	44465.09
22	HLA-A0301	KETINEEA	KE-TINEEA	0	2	1	0	0	KETINEEA	Gag_180_209	0.010	44649.25
0	HLA-A0301	TPQDLNTML	TPQDLNTML	0	0	0	0	0	TPQDLNTML	Gag_180_209	0.031	35876.13
1	HLA-A0301	PQDLNTMLN	PQDLNTMLN	0	0	0	0	0	PQDLNTMLN	Gag_180_209	0.029	36353.23
2	HLA-A0301	QDLNTMLNT	QDLNTMLNT	0	0	0	0	0	QDLNTMLNT	Gag_180_209	0.033	35061.82
3	HLA-A0301	DLNTMLNTV	DLNTMLNTV	0	0	0	0	0	DLNTMLNTV	Gag_180_209	0.056	27138.82
4	HLA-A0301	LNTMLNTVG	LNTMLNTVG	0	0	0	0	0	LNTMLNTVG	Gag_180_209	0.021	39713.52
5	HLA-A0301	NTMLNTVGG	NTMLNTVGG	0	0	0	0	0	NTMLNTVGG	Gag_180_209	0.043	31478.50
6	HLA-A0301	TMLNTVGGH	TMLNTVGGH	0	0	0	0	0	TMLNTVGGH	Gag_180_209	0.292	2129.03
7	HLA-A0301	MLNTVGGHQ	MLNTVGGHQ	0	0	0	0	0	MLNTVGGHQ	Gag_180_209	0.122	13419.03
8	HLA-A0301	LNTVGGHQA	LNTVGGHQA	0	0	0	0	0	LNTVGGHQA	Gag_180_209	0.021	39696.75
9	HLA-A0301	NTVGGHQAA	NTVGGHQAA	0	0	0	0	0	NTVGGHQAA	Gag_180_209	0.037	33383.30
10	HLA-A0301	TVGGHQAAM	TVGGHQAAM	0	0	0	0	0	TVGGHQAAM	Gag_180_209	0.078	21511.99
11	HLA-A0301	VGGHQAAMQ	VGGHQAAMQ	0	0	0	0	0	VGGHQAAMQ	Gag_180_209	0.020	40406.14
12	HLA-A0301	GGHQAAMQM	GGHQAAMQM	0	0	0	0	0	GGHQAAMQM	Gag_180_209	0.048	29872.45
13	HLA-A0301	GHQAAMQML	GHQAAMQML	0	0	0	0	0	GHQAAMQML	Gag_180_209	0.043	31303.24
14	HLA-A0301	HQAAMQMLK	HQAAMQMLK	0	0	0	0	0	HQAAMQMLK	Gag_180_209	0.681	31.42
15	HLA-A0301	QAAMQMLKE	QAAMQMLKE	0	0	0	0	0	QAAMQMLKE	Gag_180_209	0.041	32014.67
16	HLA-A0301	AAMQMLKET	AAMQMLKET	0	0	0	0	0	AAMQMLKET	Gag_180_209	0.033	35022.77
17	HLA-A0301	AMQMLKETI	AMQMLKETI	0	0	0	0	0	AMQMLKETI	Gag_180_209	0.057	26947.74
18	HLA-A0301	MQMLKETIN	MQMLKETIN	0	0	0	0	0	MQMLKETIN	Gag_180_209	0.045	30830.30
19	HLA-A0301	QMLKETINE	QMLKETINE	0	0	0	0	0	QMLKETINE	Gag_180_209	0.064	25009.20
20	HLA-A0301	MLKETINEE	MLKETINEE	0	0	0	0	0	MLKETINEE	Gag_180_209	0.051	28662.32
21	HLA-A0301	LKETINEEA	LKETINEEA	0	0	0	0	0	LKETINEEA	Gag_180_209	0.013	43256.44

Protein Gag_180_209. Allele HLA-A0301. Number of high binders 1. Number of weak binders 0. Number of peptides 45

Proteasomal Cleavage Prediction :

PAProC- Prediction Algorithm for Proteasomal Cleavage

PAProC I - prediction of cleavages by human and yeast 20S proteasomes

Our proteasome model

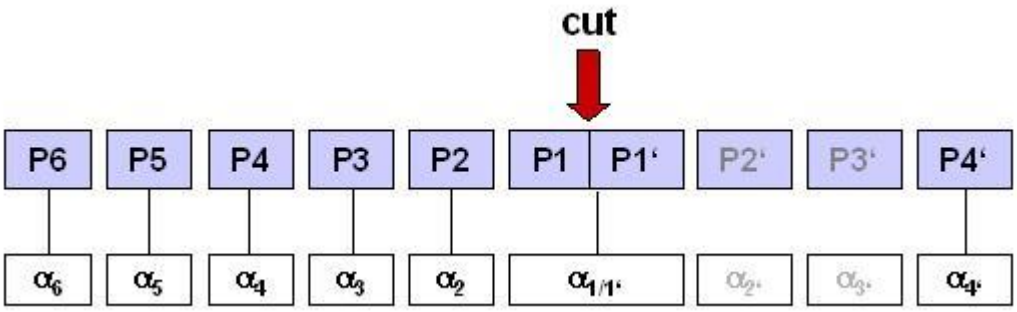
Proteasomes, major proteolytic sites in eukaryotic cells, play an important part in major histocompatibility class I (MHC I) ligand generation and thus in the regulation of specific immune responses. Their cleavage specificity is of outstanding interest for this process.

We constructed computer-based theoretical model proteasomes for the cleaving of substrate proteins by yeast and human 20S proteasomes. They were trained by an evolutionary algorithm with the experi The basic assumptions for our model are:

In determing whether to cut or not, the proteasome inspects only a small neighborhood P6 ... P1 | P1', P4' of the prospective cleavage site

A main effect results from the affinity of the pair of amino acids in the P1 and P1' positions to the active subunits in the proteasome. This effect is modeled by an affinity parameter alpha1(X1, X1'). The value - alpha1(X1, X1') could be interpreted as an affinity of the pair to the active sites of proteasome.

Each of the positions Pi, i=2,...,6 (or Pi', i=4) exerts an affinity alphai(Xi) (alphai'(Xi')) towards the prospective cut which depends on Xi (Xi') but not on the amino acids at the other positions. The affinities can be positive, negative, or zero.mental 20S proteasome cleavage data.



The model is additive: The total affinity at the position considered is:

$$\delta = \alpha_1(X_1, X_{1'}) + \sum_{i=2}^{k} \alpha_i(X_i) + \sum_{i=2}^{m} \alpha_{i'}(X_{i'})$$

A stochastic hill-climbing algorithm was used to train the network.

The affinity parameters of the model, which decide for or against cleavage, correspond with the cleavage motifs determined experimentally.

Proteasome species

Based on different sets of experimental data as learning data, we received eight different affinity parameter sets ("model proteasomes") which can be chosen:

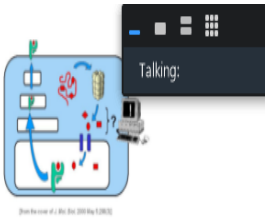
Type I: Human erythrocyte proteasome, based on cleavages in enolase

Type II: Human erythrocyte proteasome, based on cleavages in enolase and ovalbumin peptides

Type III: Human erythrocyte proteasome, based on cleavages in enolase and other ovalbumin peptides

The yeast proteasomal mutants are denoted by the missing active unit, all yeast model proteasomes are based on cleavages in enolase.

XPAPROC - Results



Short output form

The program predicts the following (**121**) proteasomal cleavages (made by **human proteasome type III**) in **Enolase** (437 amino acids):

1	AVSKVYA | RSV | YDSR | GNPTV | E | V | EL | TTEK | GVFR | SIVPSGA | ST
41	GVHE | ALEM | RDGDKSKWMGKGV | LHAV | KNV | ND | VIAPAFV | K | A | N
81	IDVK | DQKAVD | DFLISL | DGTANKSK | LG | A | NAILGV | SLAASRA
121	A | AAE | KNV | PLY | KHLADLSKSKTS | PYV | L | PVPFL | NV | LNGGSHA
161	GGAL | AL | Q | EF | MIAPTGA | KTF | AE | AL | R | IGSEVY | HNL | KSL | TKKR |
201	YGASAGNVGD | EGGVA | PNIQ | TAE | E | A | LDLIVD | AIKA | A | GHDGK
241	VKIGLDCA | SSEFF | KDGKYD | I | DF | KNPNSDKS | KW | L | TG | TQL | A | D
281	L | YH | SI | M | KR | Y | PIVSI | EDPFAED | DWEAW | SH | FF | KT | A | GIQI | V | AD |
321	DL | TVTNPKRI | ATAIEKKA | AD | ALLL | KV | NQ | I | GTL | SE | SIKA | A | Q
361	D | SF | AA | GWGV | MV | S | H | R | SGE | TED | TFI | ADLVV | GL | RTGQI | KTGAP
401	ARSE | R | LA | KLN | QLLRIE | EELGD | NA | VF | AG | E | NF | HHGDKLL

XPAPROC - Results

Long output form

The program predicts the following **(9)** proteasomal cleavages (made by **human proteasome type III**) in **pp_89** (25 amino acids):

Position	Amino acid	Cleavage prediction
1	R	non-cleavable area
2	L	non-cleavable area
3	M	non-cleavable area
4	Y	non-cleavable area
5	D	non-cleavable area
6	M	-
7	Y	++
8	P	-
9	H	

NetChop

The NetChop server produces neural network predictions for cleavage sites of the human proteasome.

NetChop has been trained on human data only, and will therefore presumably have better

performance for prediction of the cleavage sites of the human proteasome. However, since the proteasome structure is quite conserved, we believe that the server is able to produce reliable predictions for at least the other mammalian proteasomes.

This server is an update to the Netchop 2.0 server. It has been trained using a novel sequence encoding scheme, and an improved neural network training strategy. The Netchop 3.0 version has two different network methods that can be used for prediction. C-term 3.0 and 20S 3.0.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

C-term 3.0 network is trained with a database consisting of 1260 publicly available MHC class I ligands (using only C-terminal cleavage site of the ligands). 20S network is trained with *in vitro* degradation data published in Toes, et al. and Emmerich et al. C-term 3.0 network performs best in predicting the boundaries of CTL epitopes.

Instructions

In order to use the NetChop server for prediction on amino acid sequences:

1. Enter the sequence in the sequence window, or give a file name.

   The sequence must be written using the one letter amino acid code: `acdefghiklmnpqrstvwy' or `ACDEFGHIKLMNPQRSTVWY'.
   Other letters will be converted to `X' and treated as unknown amino acids.
   Other characters, such as whitespace and numbers, will simply be ignored.

2. (optional) Select prediciton method
3. Change the threshold: to increase the threshold results in better specificity, but worse sensitivity.
4. Press the "Submit sequence" button.
5. A WWW page will return the results when the prediction is ready. Response time depends on system load.

Output Format

Description

The long format output (default) consists of 5 columns:

- Residue number.
- Amino Acid
- Asigned Prediction ('S' for prediction > threshold, '.' otherwise) NOTE: the predicted cleavage site is after the assigned 'S' *i.e.* the peptide-bond on the C-terminal side of an amino acid with an assigned 'S' is cleaved.
- Predcition score
- Sequence name

When short output is chosen each input sequence will be shown with the predicted cleavage site indicated, with symbol S.

For each sequence the prediction ends with a line stating how many cleavage sites were identified.

EXAMPLE OUTPUT

Example output (long format)

```
---------------------------------------
 pos  AA  C      score      Ident
---------------------------------------
...

  74   E  .   0.107631 143B_BOVIN
  75   K  .   0.117492 143B_BOVIN
  76   K  .   0.083109 143B_BOVIN
  77   Q  .   0.557462 143B_BOVIN
  78   Q  S   0.850332 143B_BOVIN
  79   M  .   0.123313 143B_BOVIN
  80   G  .   0.344005 143B_BOVIN
...
---------------------------------------
```

Number of cleavage sites 74. Number of amino acids 245. Protein name 143B_BOVIN

Example output (short format)

```
245 143B_BOVIN
TMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQ
.S.S...S.SS.S......S.....SSS.....
```

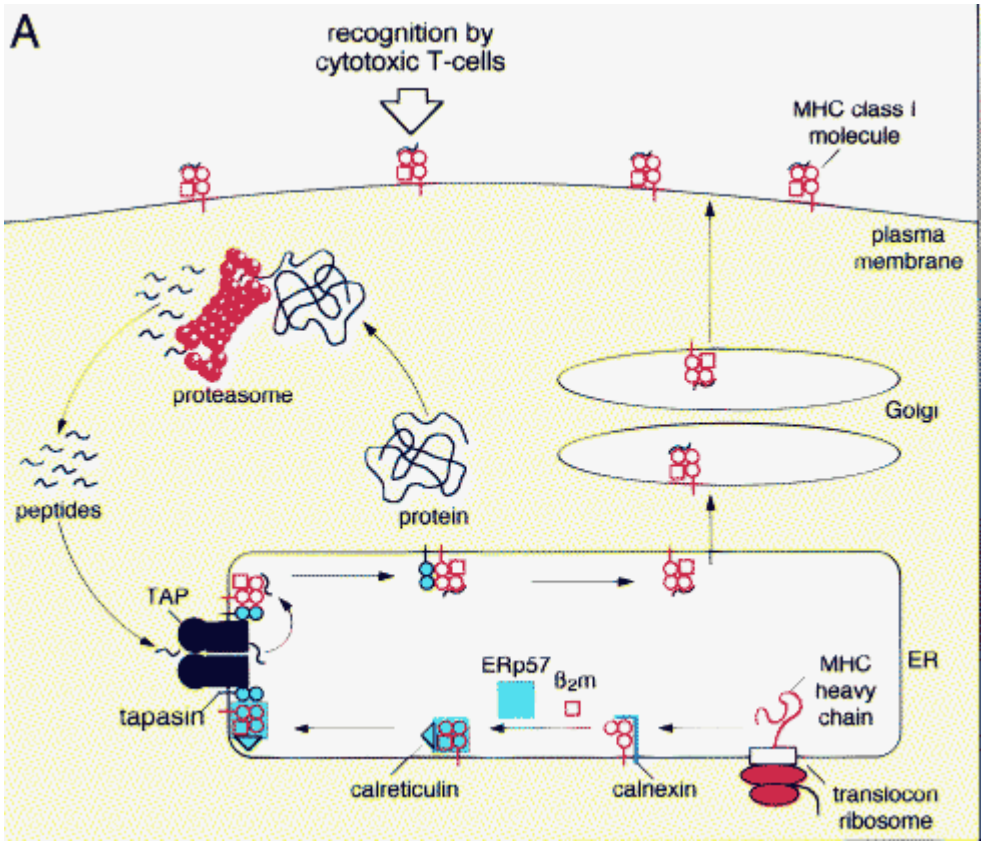
TAP Binding:TAPPred

TAPPred is an on-line service for predicting binding affinity of peptides toward the TAP transporter.The prediction of TAP binding peptides is crucial in identifying the MHC class-1 restricted T cell epitopes. The Prediction is based on cascade SVM, using sequence and properties of the the amino acids. The correlation coefficient of **0.88** was obtained by using jack-knife validation test.

General information about TAP Transporter:-

TAP is an transporter assiociated with MHC class I restricted antigen processing.The TAP is heterodimeric transporter belong to the family of ABC transporter, that uses the energy provided by ATP to transloacte the peptides across the membrane.The transporter is composed of two proteins named TAP-1 and TAP-2. The subset of these transported peptide will bind MHC class I molecules and stabilize them. These MHC-peptide complexes will be translocated on the surface of antigen presenting cells (APCs). The adducts of MHC and Peptide complexes are the ligands for T cell receptors (TCR). These complexes elicit the immune response for clearing various intracellular infections.

How does the TAP complex work is shown in the figure below.

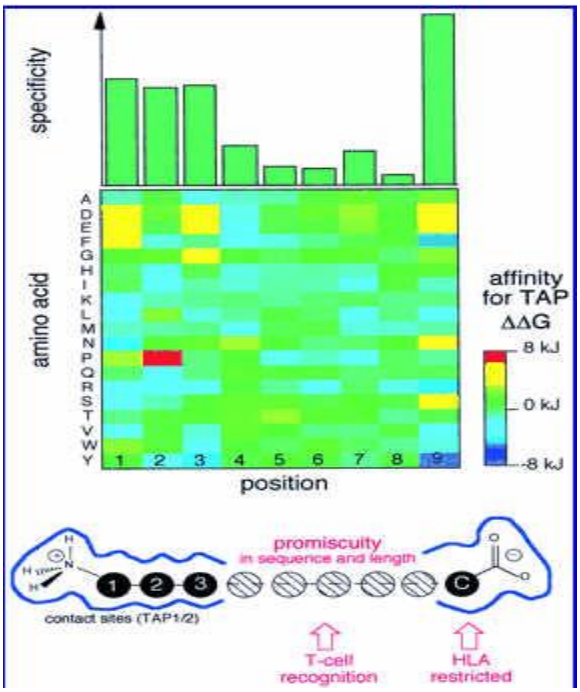


Detailed mechnaism of transport by TAP transporter

Peptide transport by TAP is a multi-step process. In a fast bimolecular association step, the peptide binds to TAP in an ATP-independent manner, followed by a slow isomerization of the TAP complex. It is suggested that this structural reorganization of the molecule triggers ATP hydrolysis and peptide translocation across the membrane. These binding steps primarily determine the selectivity of TAP. The translocation strictly requires hydrolysis of ATP, because non-hydrolyzable ATP analogs do not promote peptide transport. ATP and ADP have similar affinities for TAP; therefore, peptide translocation can be inhibited by ADP.

Peptide binding to TAP transporter

Due to extensive polymorphism of TAP transporter , distinct set of peptides will be translocated to ER. The natrure of these peptides is reflected in the nature of MHC binding peptides The selective transport of the peptides by TAP may modulate or limit the supply of the peptides to HLA class I molecules. Thus, the molecular understanding of the selectivity and specificity of TAP may contribute dramatically in the prediction of the MHC class I restricted T cell epitopes.The TAP transporter efficently bind and transport the peptides of 8-12 amino acids. It appears TAP binds peptides that are of optimal length or slighly larger then those presented by MHC class I molcules.In spite of length preference the nature of peptides has an influence on peptide selectivity. TAP from the Rat strain RT1a as well as human TAP translocate peptides with broad specificity (hydrophobic or basic amino acids at COOH terminus), whereas TAP from rat strain RT1u and TAP prefers the peptides with hydrophobic COOH termini. According to another observation TAP favours strongly hydrophobic residues in position 3 (P3) and hydrophobic and charged residues in P2, whereas aromatic and acidic residues in P1. Van Endert and Coworkers also observed that proline in position 1 and 2 have very deterious effect on binding.The TAP specificity obtained by the Peptide specificity for the TAP transporter as determined by combinatorial peptide libraries



The figure has been obtained from Lankat-Buttgereit *et al.*, 2002. **Top panel:** substrate specificity for TAP. The first three NH2-terminal amino acids and the last COOH-terminal amino acid contribute significantly to the stabilization of peptide binding to TAP. Middle panel: favored amino acids at the individual positions with negative Delta Delta G values (favored residues) are shown in blue, and positive Delta Delta G values (disfavored residues) are in red. For example, for the first position the amino acids K, N, and R are favored, and D, E, and F are disfavored. **Bottom panel:** a model of the substrate-binding pocket of TAP.

Why computaional method is required for prediction of TAP binding affinity of peptides?

The wet Lab testing of the peptides deived from the proteins is experimantally laborious and economically expensive.The Prediction methods based on the specificity of TAP transporter will complement the wet lab experiments and speed up the knowledge discoveries.on the basis of this two computational algorithms were dedeveloped in past. The algorithms are based on the machine learning technique(ANN).

We have developed a SVM based methods for the prediction of quantitative affinity of the peptides binding toward TAP. The prediction is based on complex patterns extracted from the sequence and 33 other properties of amino acids like volume, charge, aromatics residues etc. The affinity of peptide for TAP was obtained on the scale of 1-10. The correlation coefficient between the SVM prediction and measured affinity was 0.889.

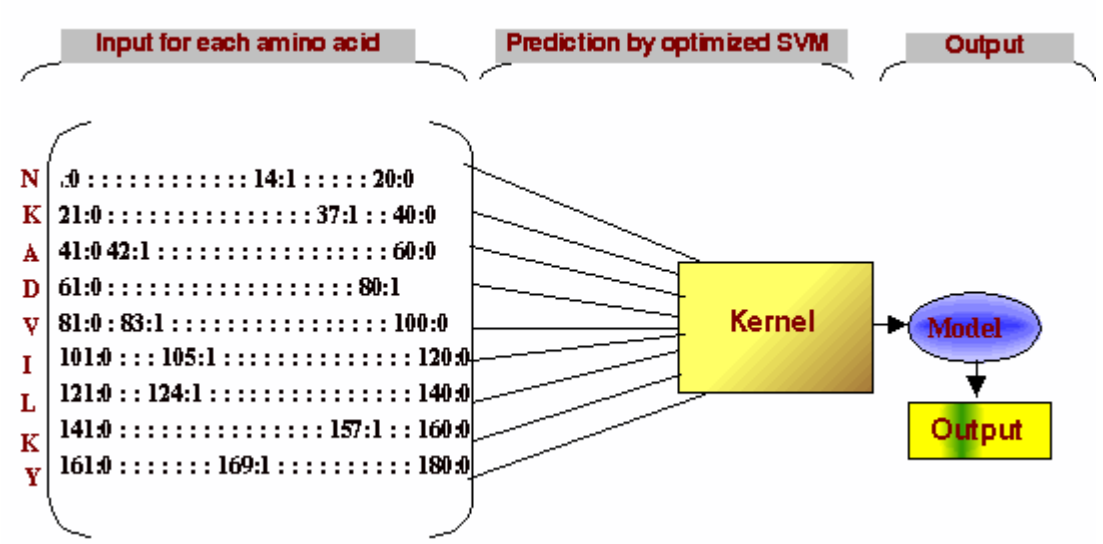
1. **Datasets for the development of prediction method:**

   The peptides dataset used in this study was kindly provided by Peter Van Endert (INSERM U580, Institut Necker, Paris France). TAP binding affinity of the peptides were expressed in term of IC50 value. The binding affinity of all peptides was tested experimentally by TAP Binding assay. The peptides have diverse binding affinity from very high (<0.03 nM) to negligible or no binding (2600 nM). All the duplicate peptides were removed from dataset. The peptides with unnatural amino acids also deleted from the dataset. The final dataset have 431 peptides with experimentally verified binding affinity. Out of 431 peptides, 179 peptides known to bind to various MHC alleles.Out of these MHC binders ,113 are present in SWISS-PROT database.

   The prediction is based on the support vector machine (SVM). Support vector machines are relatively new type of supervised machine learning that have proven to be particularly attractive to biological analysis due to their ability to handle noise and large input spaces. SVMs have been shown to perform well in multiple areas of biological analysis, including MHC binder prediction, analysis of microarray expression data and multiclass fold recognition.SVM simulation was achieved by using the **SVM_light** package. This package enable the user to define a number of parameters as well as select a choice of inbuilt kernel functions including Polynomial, RBF, Linear, Sigmoid or others. In this study the regression mode of SVM was used to model the TAP binding affinity of peptides.

Algorithm for Simple SVM:-

The simple SVM was generated on the basis of binary encoding of the sequence. Each amino acid was encoded as a 20-bit string with a unique position set at 1 and all other positions set at 0. Each peptide of 9 aa was represented by 180 inputs and a target value during the generation of the model. The target value is a real value varying from 0-10. The models were generated by using the different type of the kernels like polynomial, RBF and linear. The best model was generated by varying parameters of kernel and regulatory parameter C. The performance of the standard kernel function was evaluated by using the Jack knife testing. The performance of the kernel was determined by measuring the correlation coefficient between predicted and experimentally measured values.The overview of the final model was shown in figure below.



The correlation between the predicted and measured binding affinity reached 0.81 with the simple polynomial kernel. The performance was evaluated by using the jackknife testing. The results clearly demonstrate that polynomial kernel is more accurate; therefore it is considered as the best.The various parameters of the polynomial kernel are listed below.

Kernel::Polynomial
Regulatory parameter(C) ::5.00
Dgree of Kernel ::1.00
Correlation Coefficient::0.81

Algorithm for Cascade SVM:-

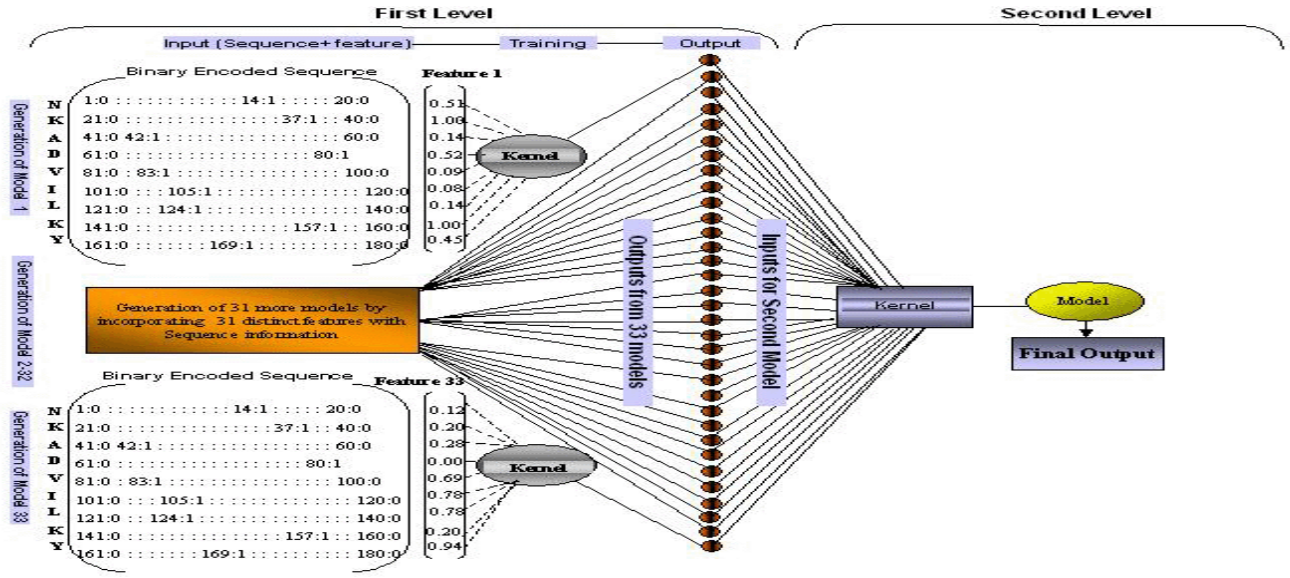
In cascade SVM, prediction is based on the sequence and features of amino acids.At first level, 33 models were generated by combining 33 features of amino acids with sequence information (one each time). At second level, final model was generated by giving the output of first level as input.

First Level:-

Models were generated on the basis of sequence and features of amino acids.The input vector for each amino acid is 21 dimensional. Among these, first twenty units of the vector stands for one type of amino acids. In order to specify particular feature of residue like charge, volume, etc, the 21st unit is added for each residue.In this manner, combining single feature of amino acids to sequence information resulted in 33 feature specific models.The overview is shown in figure below.

Second Level:-

The second model takes the outputs of the 33 models generated at first level and yields the final output on the base of these outputs. Each peptides of 9 amino acids are encode by 34 real values units, where one unit codes for the targeted value and rest 33 inputs are outputs of each peptide from 33 models generated in first level.The best model was chosen after experimenting with various types of kernels and varying their parameters. The model was fine tuned by changing the value of regulatory parameter C.



Total 33 models were generated by considering 33 features of amino acids. The analysis of the results demonstrates that none of the feature of amino acids in combination with sequence information results in significant improvement in correlation between the predicted and measure

binding affinity.Using another model of SVM, we have filtered or correlated the results of first model. The second model was fed with the output of each of 33 models generated at first level. The best result were considered were the maximum correlation between the predicted and measured binding affinity were obtained after jackknife validation testing. Using the second model, the value of correlation coefficient between predicted and measured binding affinity reached to 0.88, which is significantly higher in comparison to only sequence based prediction. The best resulted obtained at first and second level along with parameters and kernels are listed below.
First Level::-

Kernel::Polynomial
Regulatory parameter(C) ::5.00
Dgree of Kernel ::1.00
Correlation Coefficient::0.80

Second Level::-

Kernel::RBF
Regulatory parameter(C) ::30.0
Kernel Parameter (g) ::2.00
Correlation Coefficient::0.889

Results and Conculsion:-

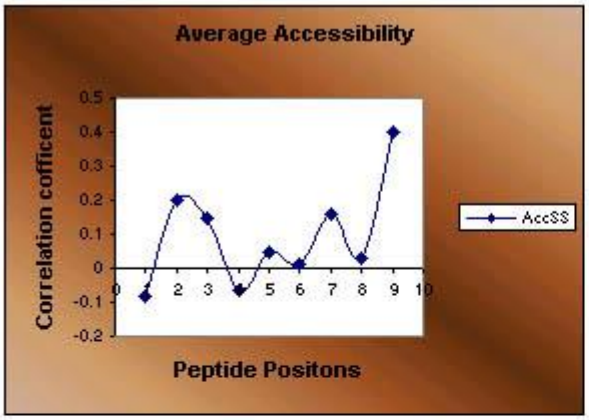
The outlines of the results obtained are shown in table below. The results clearly demonstrate that SVM outperformence the ANN in the classfication of data of TAP binding peptides.The results obtained by using the sequence based simple SVM model are better as compared to ANN based method. The correlation coffiecent of 0.732 is obatined between the measured and predicted values in previously published ANN based method.To further improve the reliability of prediction we have icoprtaed the feature information of amino acids along with sequential information.We have tried in number of ways to incoperate the fetures along with sequence information. The SVM model was generated by incorporating features of amino acids along with sequence information. The features of amino acids include 33 physiochemical properties. This results in insignificant improvement in performance of prediction method. A significant lack of improvement in the performance of prediction methods may be the result of complexity of input patterns. The SVM model generated only on the basis of features of amino acids is not able to perform comparable to only sequence-based model. The poorer performance of the features based method may be due to overlapping features of amino acids.In last we have adopted the cascade SVM based statergy for more reliable prediction. In cascase SVM the two SVM models were used. The Two models are able to predict the affinity of peptides toward TAP transporter more accurately as compared to sequentail models. The correlation coffiecent of .889 was achieved between the predicted and measured values. The outlines of the results are shown in table below.

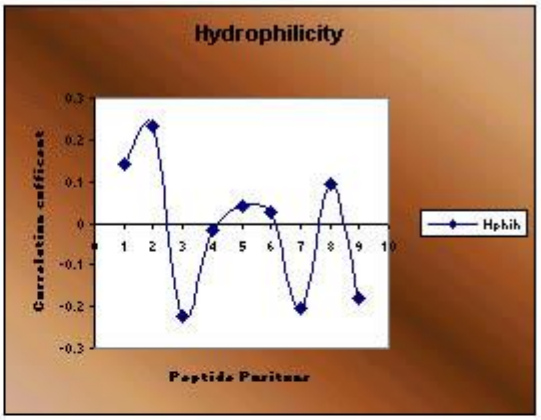
SVM Models	Polynomial Kernel		RBF Kernel	
	Parameters	Correlation coefficient	Parameters	Correlation coefficient
Only Sequence Based	C=5.00 D=1	0.812	C=15.00 G=0.005	0.795
Only Properties Based (33 Properties)	C=5.05 D=1	0.80	C=14.1 G=0.005	0.793
Sequence + Properties Based (33)	C=0.5 D=1	0.819	C=16.1 G=0.005	0.825
Cascade SVM				
First Model* (Average result of 33 models)	C=5.00 D=1	0.80	-NA-	-NA-
Final Model	C=1 D=3	0.86	C=30 G=2.0	0.88

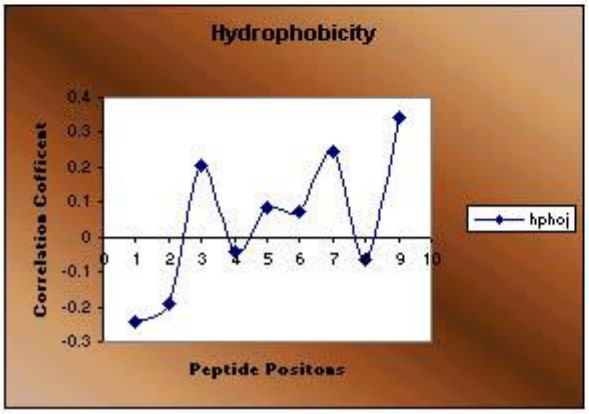
However, for more reliable prediction of TAP affinities of individual peptides, it can be envisioned to increase the predictive performance by retaining the SVM with additional data.In conclusion, human tap may skew the HLA class I associated system of antigen processing and presentation to its main task, the display of abundant of non-self proteins derived from viral or bacterial sources.

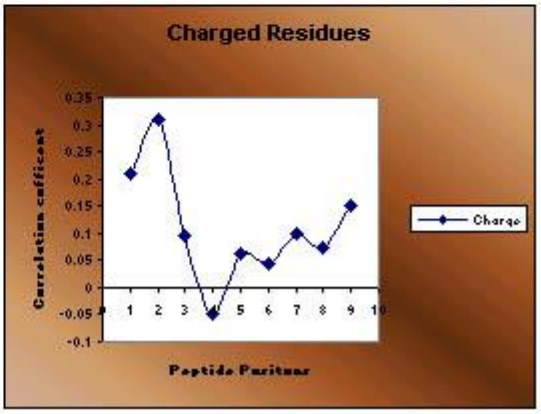
Analysis of Peptides Interacting with TAP:-

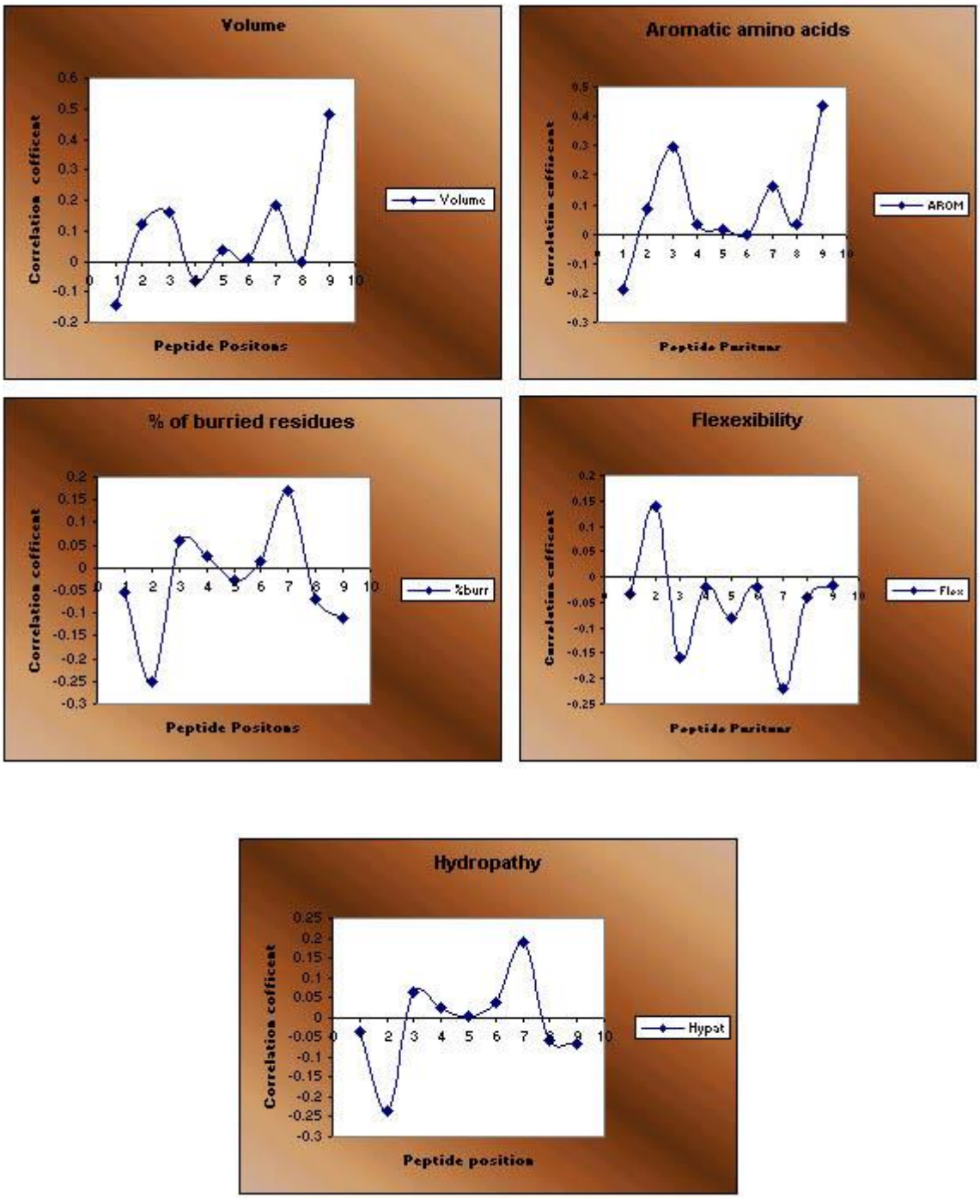
All peptides interacting with the TAP were analyzed in term of features (physical and chemical properties) of different positions (P1-P9). The TAP interacting peptides were analyzed in term of following features (Volume, Charge, aromatic, hydrophobicity, hydrophilicity, average accessibility, flexibility, hydropathy, %buried). The analysis was based on the assumption that a overrepresentation of particular property at particular position will have positive effect on affinity whereas under representation of particular property at particular position will help detrimental effect on binding. The binding affinity (IC50 value) of peptides used in analysis were expressed on the scale of 0 to 10, representing a 5-log range of normalized IC50 value from >1000 (score 0) to <0.003 (score 10) with a score increment of 1 corresponding to three fold smaller IC50 value. The values of each feature are normalized between 0 and 1. The effect of each feature for different positions of peptide is obtained by measuring correlation between feature and measured binding affinity values. The variation in each feature along the peptides (P1-P9) can be easily analyzed by plotting a graph between correlation coefficient and peptide positions.The results of the analysis are shown in graphs below.


Average Accessibility
Correlation cofficent
Peptide Positons
AccSS


Hydrophilicity
Hphih


Hydrophobicity
Correlation Cofficent
Peptide Positons
hphoj


Charged Residues
Charge



These graphs of figure 4 clearly demonstrate that three positions at N terminal and COOH terminal favors the residues with particular features. The position 1 (P1) of peptides favors the charged and hydrophilic residues, whereas the aromatic, higher volume and hydrophobic residues are not favored at P1 of peptides. The higher volume, charged, hydrophilic, accessible, flexible residues are favored at the 2nd position of the peptide. The 3rd position mostly possesses higher volume, aromatic, hydrophobic and accessible residues.The COOH terminal of the peptides prefers the higher volume, charged, aromatic, hydrophobic and accessible residues.

UNIT – V-IMMUNOINFORMATICS&COMPUTATIONAL VACCINOLOGY-SBIA5202

UNIT V IMMUNOGENETICS TO IMMUNOMICS

Biomolecular structure prediction using Immune inspired algorithm

Biomolecular Structure Prediction

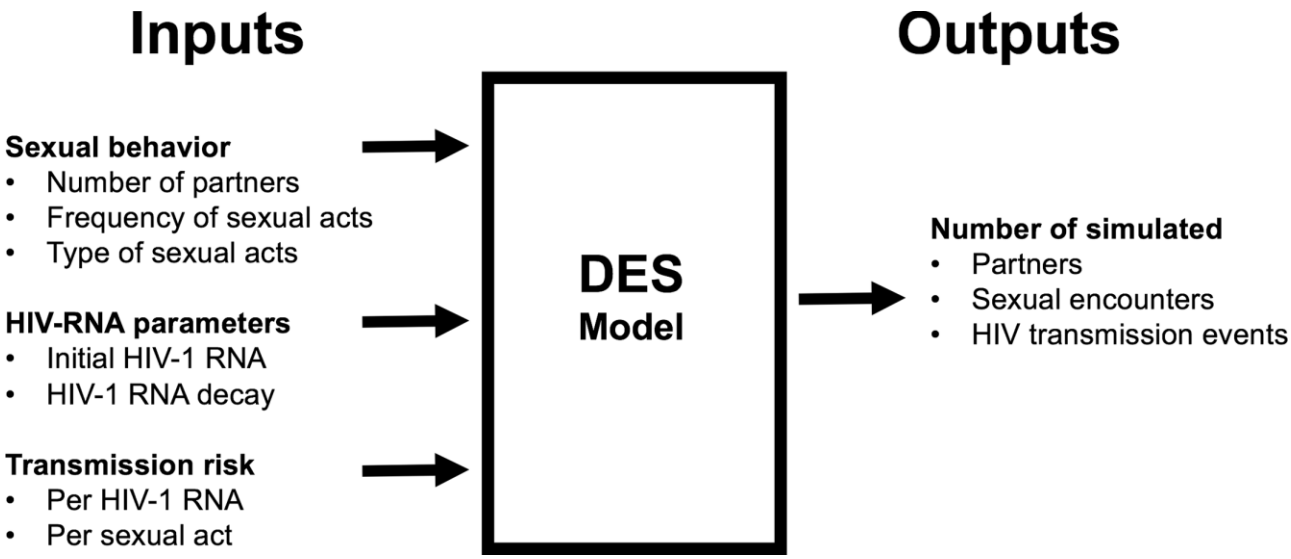
The explosion of research in molecular biology has been made possible by the fundamental discovery that hereditary information is stored and passed on in the simple,one-dimensional (ID) sequence of DNA base pairs [Watson Sz Crick 1953]. The connection between heredity and biological function is made through the transmission of this ID information, through RNA, to the protein sequence of amino acid. The information contained in this sequence is known to be sufficient to completely determine the geometrical three-dimensional (3D) structure of the protein, at least for simpler proteins which are observed to reliably refold when denatured in vitro, i.e., without the aid of any cellular machinery such as chaperones or steric constrains due to the presence of a ribosomal surface [Anfinsen 1973].

Folding to a specific structure is typically a prerequisite for a protein to function. Further understanding of the molecular description of life requires answering the deceptively simple question of how the ID sequence of amino acids in a protein chain determines its 3D folded conformation in space, or more precisely, the set of near native conformations. There are many large biological molecules, including: nucleic acids, carbohydrates, lipids and proteins. While each play a vital and interesting part in life, there is something special about proteins. Indeed, of all the components that make up life, almost all but proteins are relatively inert and are generally, the substrates thatare chopped changed by the action of proteins. In doing this, proteins do not act using some abstract bulk property, unlike lipids and carbohydrates, proteins act like individual agents that latch onto their 'objectives', the substrates, and cut and change them. Indeed, when located across a lipid membrane, they are also quite good at opening and shutting 'trapdoors'.

Discrete models for HIV

Initiation of antiretroviral therapy (ART) for HIV infection using regimens that include integrase strand transfer inhibitors (INSTIs) is associated with a faster decline in HIV-1 RNA than what is observed with regimens that are anchored by other ART drug classes. We compared the impact of ART regimens that include dolutegravir (DTG),

raltegravir (RAL), efavirenz (EFV), or darunavir/ritonavir (DRV/r), in treatment naïve men who have sex with men (MSM) on the probability of HIV-1 sexual transmission events (HIV-TE).



Structure of the model

A probabilistic approach was used to develop a discrete event simulation (DES) model using Microsoft Excel (2016) and Visual Basic for applications (VBA) to determine the number of anticipated sexual transmission events at each timepoint at and after ART initiation for each treatment scenario

Five million theoretical individuals were modeled to determine the number of secondary sexually transmitted HIV-1 infections arising from MSM initiating dolutegravir [DTG]-containing ART regimens versus infections arising from individuals starting ART containing comparator regimens (efavirenz [EFV], raltegravir [RAL], or darunavir/ritonavir [DRVr]-based ART) and versus no treatment.

The data for virologic decay was modeled based on the individual patient level data from the pivotal phase III trials of the INSTI dolutegravir (DTG): Single, Spring-2, and Flamingo. We determined the number of sexually transmitted HIV-1 infections from theoretical patients receiving ART containing DTG versus each trial comparator: the nnRTI efavirenz (EFV) in Single, the INSTI raltegravir (RAL) in Spring-2, and the PI darunavir/ritonavir (DRV/r) in Flamingo, and

versus no treatment. Each theoretical patient was cloned to obtain nine identical patients who were exposed to no therapy (in all three studies), and to ART containing DTG (all three studies), EFV (Single), RAL (Spring-2) and DRV/r (Flamingo). At the beginning of the simulation, time and transmission event counter variables were set to 0, and four attributes were randomly set for each simulated patient: the HIV-1 RNA decay curve for each simulated patient's ART treatment regimen, the number of sexual partners during the period of interest, the number of sexual encounters per partner, and the type and the timing of each sexual encounter. HIV-1 RNA for each clone was modelled to decay following a fractional polynomial regression (see below) obtained randomly from the observed decay kinetics of the relevant clinical trial and treatment regimen. For the untreated clones, we assumed that the baseline HIV-1 RNA in untreated patients remained stable during the entire observation period. Each time that a sexual encounter occurred, the probability of HIV-1 transmission to the sexual partner was modeled based on the type of sexual exposure and the HIV-1 RNA at the time of the encounter. If the partner became infected as a result of the sexual encounter, this was recorded in a counter variable, and the simulation proceeded to assess a new partner. The process was repeated for each partner until the end of the simulation time horizon when the time variable was reset to 0 to continue the simulation with the next cloned patient.

Simulation of HIV -1 Molecular evolution in response to chemokine receptors and antibodies

HIV is a highly mutable virus for which all attempts to develop a vaccine have been unsuccessful. Nevertheless, few long-infected patients develop antibodies, called broadly neutralizing antibodies (bnAbs), that have a high breadth and can neutralize multiple variants of the virus. This suggests that a universal HIV vaccine should be possible. A measure of the efficacy of a HIV vaccine is the neutralization breadth of the antibodies it generates. The breadth is defined as the fraction of viruses in the Seaman panel that are neutralized by the antibody. Experimentally the neutralization ability is measured as the half maximal inhibitory concentration of the antibody (IC_{50}). To avoid such time-consuming experimental measurements, we developed a computational approach to estimate the IC_{50} and use it to determine the antibody breadth. Given that no direct method exists for calculating IC_{50} values, we resort to a combination of atomistic modeling and machine learning. For each antibody/virus complex, an all-atoms model is built using the amino acid sequence and a known structure of a related complex. Then a series of descriptors are derived from the atomistic models,

and these are used to train a Multi-Layer Perceptron (an Artificial Neural Network) to predict the value of the IC_{50} (by regression), or if the antibody binds or not to the virus (by classification). The neural networks are trained by use of experimental IC_{50} values collected in the CATNAP database. The computed breadths obtained by regression and classification are reported and the importance of having some related information in the data set for obtaining accurate predictions is analyzed. This approach is expected to prove useful for the design of HIV bnAbs, where the computation of the potency must be accompanied by a computation of the breadth, and for evaluating the efficiency of potential vaccination schemes developed through modeling and simulation.

Integration of immune models using Petri Nets.

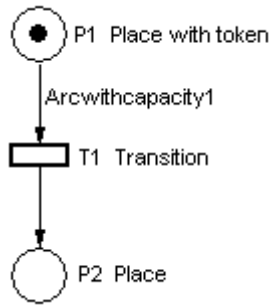
A **Petri** net is a directed bipartite graph that has two types of elements, places and transitions, depicted as white circles and rectangles, respectively. A place can contain any number of tokens, depicted as black circles. A transition is enabled if all places connected to it as inputs contain at least one token.

Petri Nets were developed originally by Carl Adam Petri [Pet62], and were the subject of his dissertation in 1962. Since then, Petri Nets and their concepts have been extended and developed, and applied in a variety of areas: Office automation, work-flows, flexible manufacturing, programming languages, protocols and networks, hardware structures, real-time systems, performance evaluation, operations research, embedded systems, defence systems, telecommunications, Internet, e-commerce and trading, railway networks, biological systems.

This introduction deals with the graphical aspect of Petri Nets for system description, not the algebra of Petri Nets. While the mathematical properties of Petri Nets are interesting and useful, the beginner will find that a good approach is to learn to model systems by constructing them graphically, aided in construction and analysis by computer software for simulation and analysis of Petri Nets.

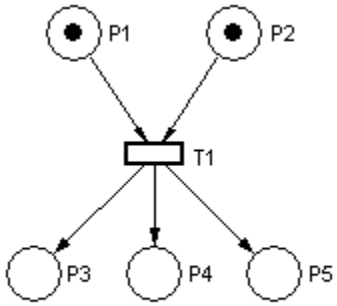
The Basics:

A Petri Net is a collection of directed arcs connecting places and transitions. Places may hold tokens. The state or marking of a net is its assignment of tokens to places. Here is a simple net containing all components of a Petri Net:

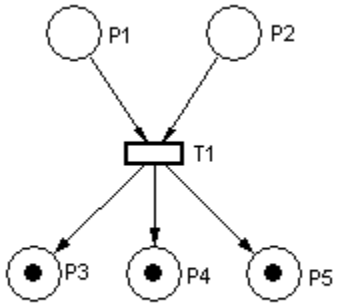


Arcs have capacity 1 by default; if other than 1, the capacity is marked on the arc. Places have infinite capacity by default, and transitions have no capacity, and cannot store tokens at all. With the rule that arcs can only connect places to transitions and vice versa, we have all we need to begin using Petri Nets. A few other features and considerations will be added as we need them.

A transition is enabled when the number of tokens in each of its input places is at least equal to the arc weight going from the place to the transition. An enabled transition may fire at any time. When fired, the tokens in the input places are moved to output places, according to arc weights and place capacities. This results in a new marking of the net, a state description of all places.

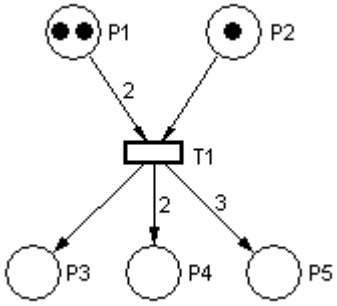


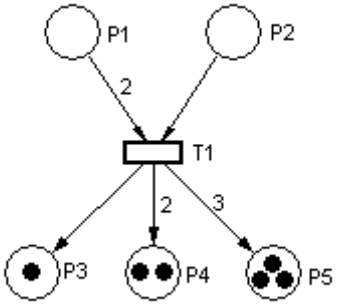
Enabled.ready to file



Firing complete

When arcs have different weights, we have what might at first seem confusing behavior. Here is a similar net, ready to fire:

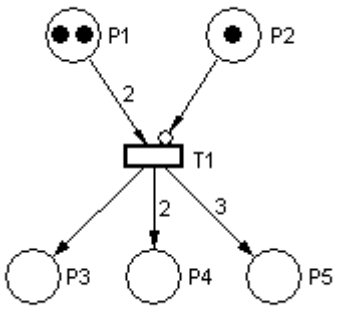




and here it is after firing:

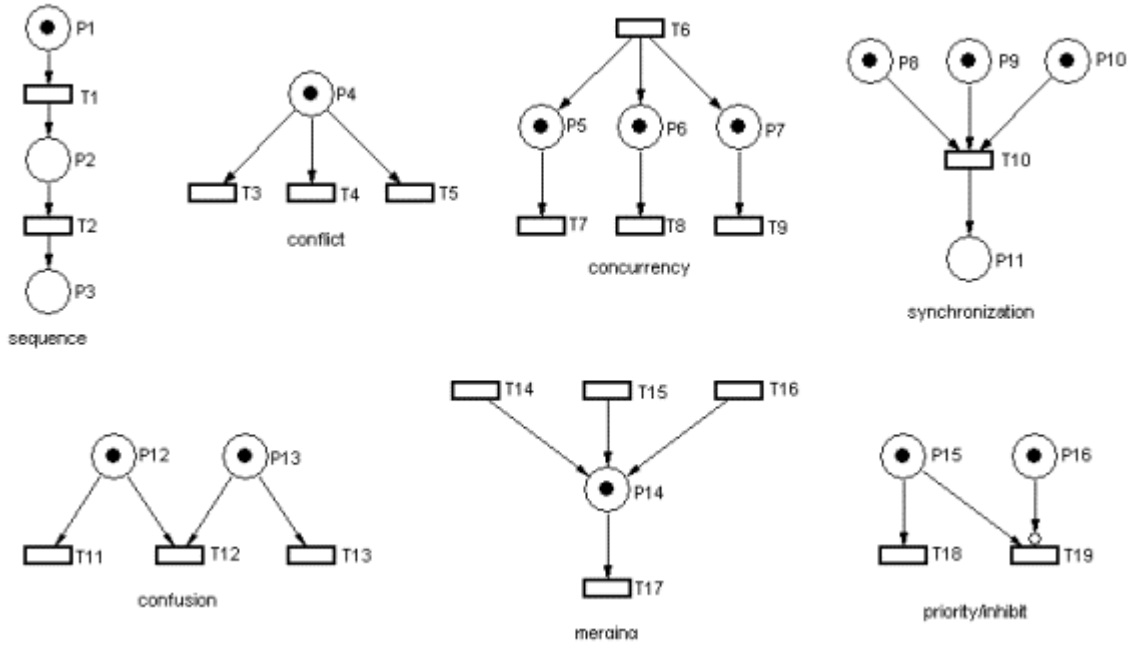
When a transition fires, it takes the tokens that enabled it from the input places; it then distributes tokens to output places according to arc weights. If the arc weights are all the same, it appears that tokens are moved across the transition. If they differ, however, it appears that tokens may disappear or be created. That, in fact, is what happens; think of the transition as removing its enabling tokens and producing output tokens according to arc weight.

A special kind of arc, the inhibitor arc, is used to reverse the logic of an input place. With an inhibitor arc, the absence of a token in the input place enables, not the presence



This transition cannot fire, because the token in P2 inhibits it.

Here is a collection of primitive structures that occur in real systems, and thus we find in Petri Nets.



Sequence is obvious - several things happen in order. Conflict is not so obvious. The token in P4 enables three transitions; but when one of them fires, the token is removed, leaving the remaining two disabled. Unless we can control the timing of firing, we don't know how this net is resolved. Concurrency, again, is obvious; many systems operate with concurrent activities, and this models it well. Synchronization is also modeled well using Petri Nets; when the processes leading into P8, P9 and P10 are finished, all three are synchronized by starting P11.

Confusion is another not so obvious construct. It is a combination of conflict and concurrency. P12 enables both T11 and T12, but if T11 fires, T12 is no longer enabled.

Merging is not quite the same as synchronization, since there is nothing requiring that the three transitions fire at the same time, or that all three fire before T17; this simply merges three parallel processes. The priority/inhibit construct uses the inhibit arc to control T19; as long as P16 has a token, T19 cannot fire.

Very sophisticated logic and control structures can be developed using these primitives.
Petri net matrix representation
Example:

Here is an example of a Petri Net model, one for the control of a metabolic pathway. Tool used: Visual Object Net++