***In-silico* analysis of host-pathogen interaction: a case study with molecular mimicry phenomenon**

# Chapter 1 Introduction

# 

# 1.1. Introduction

The host defense system is a highly complex and advanced mechanism that protects the host and eliminates infectious organisms and other intruders. It mainly helps us to avoid any type of infectious microbes, such as bacteria, virus and fungus, and it will clear them out from the body. The complete process is referred to as the immune response of the host. The immune process involves different types of cells (and their protein products), tissues and organs of the body in a highly coordinated manner.

The most important cell types which are involved in the immune process are white blood cells (WBC) also called leukocytes or leucocytes, which basically protect the body against both infectious disease and foreign invaders. Leukocytes are produced and derived from the hematopoietic stem cells in the bone marrow, found throughout the body and lymphatic system. These cells circulate via lymphatic and blood vessels between the lymphatic organs and nodes. WBC are classified in standard ways as follows: (a) on the basis of structure it can be classified as granulocytes and agranulocytes, (b) leukocytes are distinguished by cell lineage (myeloid cells or lymphoid cells). Secondly, these cells were classified on the basis of physical and functional characteristics as follows; neutrophils, eosinophils, basophils, lymphocytes, and monocytes. Furthermore, lymphocytes can be classified into B-cells, T-cells, and natural killer (NK) cells. All these cells work together to monitor the body for foreign invasion, and through blood and lymphatic vessels, they circulate between the lymphatic nodes and different organs of the body.

NK-cells are the predominant innate lymphocyte subsets that mediate antitumor and antiviral responses. B-cells are mainly involved in adaptive immunity, which produces and matures into bone marrow. Although, T-cells are produced in the bone marrow and premature cells migrate into the thymus gland for maturation, and leading to the production of mature cytotoxic and helper T cells reveals the selection process that are significant for distinguishing self from nonself. The T-cells selection criteria is quite rigorous; ~98% of the thymocytes, the precursors of T cells, die before the completion of the maturation process. Although thymic selection is remarkably effective in suppressing the immune response to self-antigens, failures do occur. Such failures result in autoimmune diseases*.*

Autoimmunity is a condition in which the immune system is instigated against self. A long heterogeneous list of diseases that range from mild to life-threatening conditions are reported to be caused by the malfunction of the immune system [(Laxminarayana 2017)](https://paperpile.com/c/a3wH5Q/sgTse). Autoimmune disease can be caused by many factors. It can be environmental or genetic or external [(Cooper et al. 1999, Jörg et al. 2016)](https://paperpile.com/c/a3wH5Q/e9Ve1+Po01p). Sometimes infectious agents may also trigger the process of autoimmunity [(Kivity et al. 2009, Sfriso et al. 2010, Wucherpfennig 2001)](https://paperpile.com/c/a3wH5Q/K19KO+DQDnT+Iuig8). This phenomenon is popularly known as molecular mimicry [(Albert & Inman 1999, Rojas et al. 2018)](https://paperpile.com/c/a3wH5Q/xt5CD+aYJP3). Microbial pathogen-induced molecular mimicry mostly occurs when antigenic determinant on one of the microbe’s proteins is similar to a determinant on one of the proteins made by the host. This results in activation of autoreactive T or B cells, which ultimately leads to autoimmunity. Though the host possesses a check mechanism to prevent elicitation of immune response against self in form of major histocompatibility complex (MHC). Hence, to generate an immune response against self, the difference between foreign and self-peptide should be strong enough, so that the self should be recognized and discriminated against by the immune system. This cross-reactivity between host and pathogen’s epitopes eventually leads to autoimmunity that may have deleterious effects onto the host [(Rojas et al. 2018)](https://paperpile.com/c/a3wH5Q/aYJP3). The disease caused by the rogue immune response is termed an “autoimmune disease”.

# 1.2. Aims and objectives

A number of studies have deciphered various prospects and aspects of molecular mimicry, but these are scattered in numerous research papers. Compilation of the available information from literature can greatly facilitate the researchers who work in this domain. At the start of my PhD work we couldn’t find any unified information resource that has a collation of all available information related to autoimmune diseases caused due to molecular mimicry. A database namely mimicDB was present that provides information about proteins or epitopes involved in host-pathogen interactions [(Ludin et al. 2011)](https://paperpile.com/c/a3wH5Q/OSJm). But mimicDB was restricted to information pertaining to only a few human parasites. Also, the mimicry candidates of mimicDB were predicted through a computational pipeline [(Ludin et al. 2011)](https://paperpile.com/c/a3wH5Q/OSJm). Therefore, I started my work with the establishment of a freely accessible database named as miPepBase **(Mi**micry **Pep**tide Data**base),** with the aim to provide a comprehensive and high quality resource of epitopes involved in molecular mimicry. All molecular mimicry based autoimmunity events of miPepBase were collected from peer-reviewed publications [(Garg et al. 2017)](https://paperpile.com/c/a3wH5Q/ElOF5).

Molecular mimicry based host and pathogen peptide interaction is reported by many viruses to disrupt or modulate host pathways to survive inside the host [(Nussinov et al. 2014)](https://paperpile.com/c/a3wH5Q/R79ms). Some of the functional machinery which viruses hijack by use of molecular mimicry are related to evasion and modulation of complement system and apoptosis, signaling pathways; control target protein levels or perturb post-translational modifications of host proteins [(Barnett & Fujinami 1992, Farris et al. 2000, Fujinami et al. 2006, Rosen et al. 1995)](https://paperpile.com/c/a3wH5Q/xUgMP+MlDsU+EQ17z+QbIF1). The proteins, which are involved in molecular mimicry in viruses, are mostly disordered and they manipulate the host cellular mechanism by harboring short linear motifs (SLiMs) and molecular recognition features (MoRFs) on their protein [(Hraber et al. 2020)](https://paperpile.com/c/a3wH5Q/dy5XJ). The mimicry motifs in viruses are not continuous and are located in the disordered regions of proteins [(Duro et al. 2015)](https://paperpile.com/c/a3wH5Q/vZh5g). In comparison to viruses, very little work has been done to understand the finer details of interaction between host and bacterial memitopes. In this thesis, we also analysed structural and functional properties of bacterial and viral host/pathogen mimicry proteins. We also explained the benefit of molecular mimicry in terms of immunological consequences in microbial pathogenesis.

The conventional therapy for autoimmune diseases has been the usage of immune-suppressants or immune-modulators that treat symptoms rather than the etiology and/or the causative mechanism(s) [(Rosenblum et al. 2012)](https://paperpile.com/c/a3wH5Q/92SP7). Molecular mimicry can cause several autoimmune diseases along with the disease caused by the infectious agents [(Rojas et al. 2018)](https://paperpile.com/c/a3wH5Q/aYJP3). Hence, the process of molecular mimicry can be used as a stepping-stone to understand the initial interaction between the infectious agent as well as recognizing the self-determinant, understanding the pathogenic mechanism(s) involved, and designing strategies for the treatment and prevention of autoimmune disorders. Though, 60–70% of the patients initially respond to immunosuppression, in many cases the patients show subsequent clinical remission or relapse of the autoimmune disease [(van der Kooij et al. 2007)](https://paperpile.com/c/a3wH5Q/37qW9). Henceforth, we designed a novel strategy for the treatment and prevention of autoimmune disorders. In which, we used molecular mimicry proteins and its interaction partners as a potential target to block the production of molecular mimicry proteins, which will ultimately lead to killing of pathogens.

In summary, the overall objectives of this thesis are as follows:

i. Creation of a data repository of experimentally verified molecular mimicry phenomena that leads to autoimmune diseases.

ii. Analysis of structural, functional and immunological characteristics of bacterial-, viral- and host-mimicry proteins and peptides.

iii. Exploration of proteins involved in molecular mimicry as novel drug targets. A case study with (a) *Mycobacterium tuberculosis* (b) *Mycobacterium avium* subsp. *paratuberculosis*

iv. Functional analysis of *M. tuberculosis* mimicry phenomena.

# 1.3. Structure of the thesis

This thesis is organized into 8 chapters. Chapter 1 is Introduction to the general framework of the work presented in this thesis. Chapter 2 is Review of Literature in which a basic description of molecular mimicry induced autoimmune diseases are discussed. Chapter 3 describes a database named as miPepBase - containing experimentally validated information related to autoimmune diseases caused due to molecular mimicry. Further, evolutionary, structural, immunological and functional characteristics of bacterial-, viral- and host-mimicry proteins, are described in Chapter 4.

We developed a therapeutic schema, which can be applied to repurposing known drugs and/or discovery of novel therapeutics against pathogenic bacteria, which exhibit molecular mimicry with the host’s proteins. In the chapter 5, we have used *Mycobacterium* *tuberculosis* (Mtb) as a model organism and using our strategy we were able to found four drugs viz. DB08185, DB00759, DB01930 and DB07349 that might be useful in treatment of *M*. t*uberculosis*.

In Chapter 6, using our proposed approach, we found eight DrugBank molecules, which might prove useful for treating three *Mycobacterium* *avium* subsp. *paratuberculosis* (MAP)-associated autoimmune diseases, type 1 diabetes, Crohn’s disease, and multiple sclerosis. Moreover, the drug molecules identified during our analysis are either FDA-approved drugs or experimental drugs with proven efficacy. Hence, these can be easily incorporated in clinical studies or tested *in vitro* for assessing their suitability in treating MAP-associated autoimmune diseases.

In chapter 7, we explained how the *M. tuberculosis* mimicry and its interacting proteins export via exosomes, and hijacking the host’s system to maintain the dormant phase. Finally, in Chapter 8, the key issues and significant outcome of the present work is summarized, which would be helpful for the treatment and prevention of autoimmune disorders.

# 

# Chapter 2 Review of Literature

# 2.1. Innate and adaptive immunity

The immune system protects organisms from infection with two major defenses mechanisms, with increasing specificity. (i) the innate, provides the first line of defenses against infection. (ii) the adaptive, specialized immune system. Both systems do not operate independently, instead they function in a highly interactive and coordinated manner, producing a combined response more effective than either branch could produce by itself. Further, both states of immunity required several soluble substances found in blood and other body fluids, which belongs to the humoral defense mechanism. Henceforth, the innate and the adaptive immune system use both cellular and humoral defense strategies.

## 2.1.1. Innate immune system: Fast and broadly effective

Innate immune response is the most rapid and evolutionary conserved arm of the immune system [(Chaplin 2003)](https://paperpile.com/c/a3wH5Q/M39hT). Many innate components are present before the pathogen encounter and constitute a set of disease resistance mechanisms which are not specialized for particular microbes but contain molecular and cellular components that can identify a particular class of molecules occurring in frequently encountered pathogens.

The innate defense consists of several elements:

* External barrier, include skin and all mucous membranes
* Acidity of stomach content and perspiration
* Phagocytosis - conducted by a group of specialized cells such as blood monocytes, neutrophils and tissue macrophages. They can recognize and neutralize invaders based on common molecular surface markers.
* Enzymes such as lysozyme - a hydrolytic enzyme found in mucous secretion and in tears, attack the peptidoglycan layer of bacteria cell wall.

# 2.1.2. The adaptive immune system: Precision and a long memory

When the body’s first line of defense fails to completely eliminate the pathogen, then the adaptive immune system comes into picture, for complete eradication of invaders. As it takes longer time to recognize (approx. four to seven days), but selectively eliminates specific foreign microbes. But, defense response will be more efficient and faster, if the body comes into contact with already known antigens, because it has immunological memory [(Ahmed 1992)](https://paperpile.com/c/a3wH5Q/x9X8D).

Adaptive immunity exhibits four characteristics attributes: (a) Antigenic specificity (b) Diversity (c) Immunological memory (d) Self-nonself recognition.

An adaptive immune response has various parts, activation of each part is depending upon the pathogen location into the host body. For instance, extracellular pathogens have presence of antibodies into blood, and if a pathogen is inside the tissue, a cell-mediated immune response is necessary [(Sfriso et al. 2010)](https://paperpile.com/c/a3wH5Q/Iuig8).

These parts of the adaptive defense include:

1. T lymphocytes
2. B lymphocytes
3. Antibodies as soluble proteins in the blood
4. Cytokines in the blood and tissue as hormone-like messenger substances

# 2.1.2.1. T lymphocytes or T cells

T lymphocytes are the main component of the adaptive immune system, which are involved in a particular defenses mechanism. T cells are produced in the bone marrow and premature cells migrate into the thymus gland for maturation, mature T cells are able to differentiate between self and non-self antigens. In the thymus gland, T cells develop particular surface receptors which are capable of recognizing and binding pathogens. The interaction between the T cell surface receptors and pathogens stimulates the T cells for faster division and meanwhile they activate other defenses reactions. They are capable of eliminating pathogens from the body. During the defense reaction T cells are evolved into some specialized cells. These include: T helper cells, T killer cells or cytotoxic cells, memory T cells and regulatory T cells [(Adam et al. 1998)](https://paperpile.com/c/a3wH5Q/X4JIu).

# 2.1.2.2. B lymphocytes or B-cells

B lymphocytes are the other important cells that are involved in adaptive immunity. B-cells, production and maturation both occur into bone marrow. Mature B-cells produced antibodies which are in the blood as soluble proteins are specific for particular pathogens.

To activate T- and B-cells of the adaptive immune system either direct binding to the antigen presenting cells (APC) or interaction with different types of messenger molecules like cytokines are required.

# 2.2. “Molecular mimicry”: an evolving concept

In 1964, Damian first proposed the term “molecular mimicry” that refers to the sequence or structural similarity between pathogens and their hosts, which assists microbes to avoid the host immune response [(Damian 1964)](https://paperpile.com/c/a3wH5Q/gmerF). Before introducing this term, Kaplan *et al.* provide experimental evidence of immune cross-reactivity in a rheumatic fever patient by examining the sera reaction of rabbits immunized with group A streptococcal cells to human heart tissue [(Kaplan & Meyeserian 1962)](https://paperpile.com/c/a3wH5Q/rDKkO). Though, during that time their structural homology information was unavailable. Two years later, it was found that membrane structures in group A streptococcus shared structures with mammalian muscle [(Zabriskie & Freimer 1966)](https://paperpile.com/c/a3wH5Q/j0QVL). Sometimes the structural similarity between host and pathogen may provide immune tolerance to the pathogenic organisms instead of immunogenic response activation, for e.g. structural similarity between the antigenic determinants of parasite and antigenic structures of human leads a parasitemia response (i.e. immune tolerance) in the host [(Damian 1964)](https://paperpile.com/c/a3wH5Q/gmerF). Although the concept of homology is conflicting in terms of immunogenic activation and tolerance, it suggests that these hypotheses could work together, and apart from homology some other factors are also responsible for triggering autoimmunity [(Hardtke-Wolenski et al. 2017, Ma et al. 2017, Wu et al. 2018)](https://paperpile.com/c/a3wH5Q/Fom9p+MN3N4+dJ2oD).

Though, clonal deletion prevents autoimmune response in the host. During clonal deletion, highly self-reactive lymphocytes are destroyed and prevent immune response against self-antigens [(Rose 2015)](https://paperpile.com/c/a3wH5Q/vX4ho). Sometimes, a pathogen which shows molecular mimicry took advantage over non-mimicking to avoid the host's immune response [(Drayman et al. 2013)](https://paperpile.com/c/a3wH5Q/mpgRS).

There are numerous epidemiological and experimental evidences that suggest the role of infectious diseases in autoimmunity via molecular mimicry and cross-reactivity [(Guarneri 2013)](https://paperpile.com/c/a3wH5Q/OwHVO). As of now it is not a very critically studied process of microbial pathogenesis. But slowly the role of molecular mimicry in microbial pathogenesis has been unfolding and it is implicated in a number of pathological conditions. In 1983, Fujinami *et al* observed the cross-reactivity between the murine antibodies of measles virus/herpes simplex virus (HSV) and human cells [(Fujinami et al. 1983)](https://paperpile.com/c/a3wH5Q/LBBq9). In line, they also observed that myelin basic protein (MBP) encephalitogenic peptide shares homology with the hepatitis B virus polymerase (HBVP) that results in an autoimmune disorder named as encephalomyelitis [(Fujinami & Oldstone 1985)](https://paperpile.com/c/a3wH5Q/jgXPb)

Furthermore, in nearly half of tuberculosis (TB) patients, autoantibodies responsible for wegener's granulomatosis and systemic lupus erythematosus and many others are being observed. It has been proposed by Elkington, *et al*. (2016) that TB “tricks” the immune system into attacking the lungs, enabling the bacteria to become more infectious [(Elkington et al. 2016)](https://paperpile.com/c/a3wH5Q/HZU7I).

# 2.2.1. Type of molecular mimicry

Molecular mimicry can be achieved at four different levels:

**(1) Sequence and structure mimicry**

When sequence and structure of a host protein is hijacked by pathogen protein, it will give rise to orthologous proteins across the pathogen species. Although, to adapt to a similar structure, pathogens used an evolutionary strategy named as horizontal gene transfer (also known as lateral gene transfer) [(Guven-Maiorov et al. 2016)](https://paperpile.com/c/a3wH5Q/gffNr).

An example for sequence and structure mimicry type could be the Toll/Interleukin-1 receptor (TIR) domain. Although, any immune system receptors comprised TIR domain, for e.g. Toll-like receptors (TLRs), interleukin-1 receptor (IL-1R), and downstream effectors (such as Mal, MyD88, TRAM and TRIF) [(Guven Maiorov et al. 2013, Guven-Maiorov et al. 2015)](https://paperpile.com/c/a3wH5Q/dHDA+tMPO). Further, pathogens also express TIR domain-containing proteins (Tcps) to interfere with TLR and IL-1R signaling [(Chan et al. 2009, Cirl et al. 2008)](https://paperpile.com/c/a3wH5Q/hY93+athA). For instance *Escherichia coli*, *Brucella melitensis* and vaccinia virus which encode/secrete TcpC, TcpB [(Cirl et al. 2008)](https://paperpile.com/c/a3wH5Q/hY93), and A46R [(Janeway & Medzhitov 2000)](https://paperpile.com/c/a3wH5Q/gOVW)proteins, respectively that disrupt the host’s immune signaling pathways. These bacterial TcpC and TcpB can directly bind to MyD88 that decreases the host pro-inflammatory cytokine production and promotes bacterial survival [(Cirl et al. 2008)](https://paperpile.com/c/a3wH5Q/hY93).

**(b) Structural mimicry without (or with very low) sequence similarity**

Structural similarities between the microbes are responsible for the hijacking host’s proteins. An example of complete structure conservation with very low sequence similarity is viral chemokine vMIP-II of Kaposi’s sarcoma-associated herpesvirus [(Qin et al. 2015)](https://paperpile.com/c/a3wH5Q/nzWN) and chemokine receptor US28 of human cytomegalovirus [(Burg et al. 2015)](https://paperpile.com/c/a3wH5Q/JnGM)These viral proteins have very low sequence similarity with host proteins *viz.* vMIP-II is ~33% similar to the human chemokine CX3CL1, and US28 is ~29% similar to the human chemokine receptor CXCR4 in terms of structure. The structural similarity between the surface receptors is not only advantageous for pathway inhibition/activation, but also allows pathogens to anchor to host surfaces and facilitate nutrient uptake by the host [(Finlay & McFadden 2006)](https://paperpile.com/c/a3wH5Q/DMXqx)

**(c) Motif mimicry**

Some pathogens have homologs of short amino acid sequences instead of the whole proteins or domains, known as motif mimicry [(Davey et al. 2011, Hagai et al. 2014)](https://paperpile.com/c/a3wH5Q/JLIZ+EpT9z). Short linear motifs (SLIMs) are present within (intra)- and across (inter)-species. SLIMs are generally composed of 3–10 residues, and were proposed to have important roles in pathway modification. They were proposed to have ‘evolutionary plasticity’; that is, changes of few residues in the protein can rewire pathways thereby adapting cell signaling [(Davey et al. 2011)](https://paperpile.com/c/a3wH5Q/EpT9z). Many viruses possess more than one SLIM in their genome, enabling them to interfere with more than one host interaction by competitively displacing the host protein partner [(Davey et al. 2011)](https://paperpile.com/c/a3wH5Q/EpT9z).

One example for motif mimicry is the WxxxE motif in many bacterial guanine nucleotide exchange factors (GEFs), such as Map and EspM2 of *E. coli* and SifA of *Salmonella* **54,55**. Like endogenous GEFs, these pathogenic GEFs activate the GTPases in the host. Although not located at the catalytic site, the presence of the WxxxE motif is critical for the GEFs’ interactions with the GTPases. SopE of *Salmonella* does not possess a WxxxE motif, but it still folds into a structure, which is very similar to those effectors that do. It has Y and T residues, which correspond to W and E in the motif. Thus, in spite of the different residues, it conserves the chemical properties at the corresponding sites.

**(d) Interface mimicry**

Interface structure similarity is the most common type of molecular mimicry. For host-pathogen protein–protein interaction there is no need for extensive sequence or structure conservation. Sometimes the overall structures of the proteins are distinct, but still they use similar interface architectures to interact with their partners, which suggest that these recurring architectures are favorable scaffolds [(Keskin & Nussinov 2005)](https://paperpile.com/c/a3wH5Q/AZT1). The interface mimicry assists pathogen evasion[(Yamada et al. 2015)](https://paperpile.com/c/a3wH5Q/miad)and also supports many cellular events that take place through competitive binding [(Franzosa & Xia 2011, Franzosa et al. 2012)](https://paperpile.com/c/a3wH5Q/gjiQ+VjOj) The example for interface mimicry is human fibronectin and the invasin protein of *Yersinia* bacteria, both of which bind to human integrin in a similar way (Stebbins & Galán 2001, Zur Hausen 2009). Despite lack of overall structural and sequence homology, they have similar chemical properties at the integrin-binding site.

# 2.2.2. Molecular mimicry and cross-reactivity: what is necessary?

The activation of T and B-cells against self-antigen is the major consequence of any autoimmune diseases, wherein the former lymphocytes play a significant role in ‘T-cell mediated autoimmune diseases’ [(Bertsias et al. 2010, Gregersen et al. 1987, Mathis et al. 2001)](https://paperpile.com/c/a3wH5Q/BcEsg+lvOl7+doQiB). Earlier, it was assumed that due to highly specific recognition – T cells show low cross-reactivity with infectious diseases [(Cusick et al. 2012)](https://paperpile.com/c/a3wH5Q/WvBKB). Later some researchers show that only a very small portion of an antigen is being recognized by T-cells receptors (TCR) and shows MHC restriction – interaction between specific TCR and MHC bearing peptides. In this context, 8-10 amino acids are presented by MHC class I to CD4+ T-cells, and 14-18 amino acids are presented by MHC class II to CD8+ T-cells [(Reay et al. 1994, Sinigaglia & Hammer 1994, Wucherpfennig et al. 1994)](https://paperpile.com/c/a3wH5Q/wSgSo+6KL82+Izr8G). In these short antigenic peptides, there are some anchor residues that are meant to bind specific pockets on the MHC molecules, resulting in some specificity of interactions with MHC [(Harbige et al. 2017, Paun et al. 2016, Vatti et al. 2017)](https://paperpile.com/c/a3wH5Q/FhP55+bskJY+RP1Xg). Also, there is a certain degree of plasticity in the other residues, so that different peptides or chemical xenobiotics can bind to single MHC molecules with certain specificity, even though some peptides can bind to more more than one MHC, a phenomenon known as “polyspecificity” [(Wucherpfennig et al. 2007)](https://paperpile.com/c/a3wH5Q/2N5Si).

In a case study using mice – as host and rat insulin promoter-lymphocytic choriomeningitis virus (RIP-LCMV) – as pathogen, it has been shown that heterologous sequential viral infections can increase cross-reactive T cells in the targeted organ above the disease initiating threshold, leads to major tissue injury and in this case rapid development of diabetes, an autoimmune disease [(Christen et al. 2004)](https://paperpile.com/c/a3wH5Q/5UUv5). The cross-reactivity between the host-pathogen epitopes can increase but not initiate autoimmune diseases in the host. Interestingly, this observation suggested that in humans, the combined effect of some immunologically cross-reactive viruses are the main reason for boosting the autoimmune disease(s).

## 2.3. The initiation of autoimmunity: lighting the match

The rate of morbidity and mortality in the human population increases due to autoimmune diseases (AD). The immune system is mostly able to differentiate between self and non-self antigens, if not, it activates autoimmune response into the body. There are two types of autoimmune diseases: **tissue-specific** type where antigen targeting is single tissue-specific and **systematic** type in which more than one tissue and ubiquitously expressed antigens are targeted. It is difficult to pinpoint the single factor that causes autoimmune diseases, due to delayed symptoms well after the abnormal reaction begins [(Fridkis-Hareli 2008)](https://paperpile.com/c/a3wH5Q/TmBcm). Apart from humans, some animals also show an autoimmune response. Although, many groups of researchers are using model organisms to find-out early stage diagnosis, to prevent poor prognosis [(Konforte et al. 2012)](https://paperpile.com/c/a3wH5Q/0FKzY).

The autoimmune disease is a complex mechanism, which can be triggered either by numerous infectious agents (virus and bacteria) or molecular and cellular pathways and events. Furthermore, the sequence or structure homology between infectious agents and self-antigen lead to proinflammatory response into the host, termed as molecular mimicry (cross-reactivity) and it is one of the causes of autoimmunity [(Cusick et al. 2012)](https://paperpile.com/c/a3wH5Q/WvBKB). This cross-reactivity between host’s and pathogen’s epitopes can have deleterious or protective effects onto the former.

# 2.3.1. Type of autoimmune diseases

There are more than 100 autoimmune diseases reported till date. But the most common reports autoimmune diseases are caused by ‘linear molecular mimicry’ (explained in section 3.1.1) because it is easiest to study. Some most commonly found diseases are multiple sclerosis [(Nielsen et al. 2007)](https://paperpile.com/c/a3wH5Q/lJ5nD), Guillain-Barré syndrome [(Sheikh et al. 1998)](https://paperpile.com/c/a3wH5Q/n3ktS), Type 1 diabetes [(Coppieters et al. 2012)](https://paperpile.com/c/a3wH5Q/AW4dN), Rheumatoid arthritis [(Ebringer & Rashid 2009)](https://paperpile.com/c/a3wH5Q/5MZiN), Systemic lupus erythematosus [(Poole et al. 2006)](https://paperpile.com/c/a3wH5Q/Ar7Lo), Sjögren's syndrome [(Igoe & Scofield 2013)](https://paperpile.com/c/a3wH5Q/CZgTm), Systemic sclerosis [(Grossman et al. 2011)](https://paperpile.com/c/a3wH5Q/SXpMn), Autoimmune thyroid disease [(Benvenga & Guarneri 2016)](https://paperpile.com/c/a3wH5Q/x0vPM), Autoimmune hepatitis [(Christen & Hintermann 2018)](https://paperpile.com/c/a3wH5Q/x1vi7) and primary biliary cholangitis [(Van de Water et al. 1993)](https://paperpile.com/c/a3wH5Q/uc4oX)

Interestingly, many mycobacterial antigens have been associated with autoimmune diseases. This prompted us to in depth study of mycobacteria that share sequence similarity with host antigens, and elicit T cell autoimmune reactions. Some have been discussed in detail below:

* ***Multiple sclerosis (MS)* :** MS is a chronic disease of the central nervous system in which myelin covering around the nerve fibers is significantly reduced or disappears completely. Since nerve fibers cannot efficiently conduct the electrical impulses in absence of myelin covering, hence the electrical impulses received from the brain do not flow smoothly to the target nerve. Due to this, the muscle movement becomes very erratic. Although the exact cause of MS is not known, it is believed to be a multifactorial disease caused by autoimmune processes [(Libbey et al. 2007)](https://paperpile.com/c/a3wH5Q/nODs2). Multiple sclerosis is one of the leading disease caused due to molecular mimicry. Many clinical evidence suggest that an infectious agent might be responsible for breaking tolerance and elicitation of autoimmune response against myelin proteins. Indeed, it was observed that some infectious viruses and bacteria *viz.* endogenous retrovirus, Epstein Barr virus (EBV), *Chlamydia pneumoniae*, *Helicobacter pylori*, and *Mycobacteria spp*., might have a role in the MS [(Cossu et al. 2018)](https://paperpile.com/c/a3wH5Q/B74kP). According to some reports, in comparison to healthy individuals, MS patients have enhanced proliferation of lymphocytes against *Mycobacterium tuberculosis* and *Mycobacterium leprae* - recombinant heat shock proteins 65 and 70 (HSP60 and HSP70) proteins. Some researchers also observed that in Sardinian MS patients there was an increase in circulating antibodies against *Mycobacterium avium* subsp. *paratuberculosis* (MAP) HSP70 [(Cossu et al. 2017)](https://paperpile.com/c/a3wH5Q/38sGN).
* ***Type 1 diabetes mellitus (T1DM):*** It is a second most common childhood chronic disease, in which T-cells penetration destroys the insulin producing beta cells of the pancreas. T1DM can be broadly classified into two types: (i) **type 1A -** patientshave antibodies against host-proteins such as glutamic acid decarboxylase 65 (GAD65), insulin, insulinoma associated proteins (IA-2) and heat shock protein 60 (hsp60), and (ii) **type 1B -** a few cases was reported for type 1B, has no known cause [(Rani et al. 2010)](https://paperpile.com/c/a3wH5Q/2MiS6). Additionally, GAD is a 65 Kda enzyme that catalyzes the α-decarboxylation reaction of L-glutamic acid to synthesis of gamma-amino butyric acid (GABA). GAD65 is a major autoantigens that is found in many tissues and participates in TIDM pathogenesis. Henceforth, GAD65 antibodies are more commonly used for diagnosis purposes. Although, high blood sugar level (hyperglycemia) leads to long lasting side effects on the body such as stroke, cardiovascular ailments and diabetic nephropathy/neuropathy/retinopathy. Currently, insulin replacement therapy is the only symptomatic aid for T1DM. Some experimental evidence suggests that role of *Mycobacterium avium* subsp. *paratuberculosis* in triggering T1DM. For instance, that molecular mimicry between human GAD65 and MAP Hsp65 elicits an autoimmune reaction targeting beta cells in pancreatic islets results in T1DM and insulin deficiency [(Dow 2012)](https://paperpile.com/c/a3wH5Q/dfBKf).
* ***Leprosy*:** In leprosy, autoimmune response generated against host’s components, primarily causing nerve damage. Some clinical reports suggested that in leprosy patients there is an increase in autoantibodies level and lympho-proliferative response against myelin basic protein (MBP). Further, cross-reactivity between *Mycobacterium leprae* proteins Lysyl-tRNA synthetase/50S ribosomal protein L2 and MBP (myelin A1 protein) of host caused nerve damage in leprosy [(Singh et al. 2015)](https://paperpile.com/c/a3wH5Q/pKnGV).
* ***Crohn’s disease*:** an immune-mediated inflammatory bowel disease (IBD), with unclear cause. CD shows similarity with intestinal tuberculosis and Johne’s disease. Mostly in ruminants and primates, CD is known to be caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP). Some studies prove that MAP exhibits ‘epitope mimicry’ with the human intestinal proteins, which elicits host autoreactive T- or B-cells leading to CD. In this support, comparative analysis between CD patients and healthy control shows that increased levels of antibody against specific MAP antigens and detection of more viable MAP in peripheral blood and intestinal tissue in CD patients [(McNees et al. 2015)](https://paperpile.com/c/a3wH5Q/sm84x). Reportedly, in CD patients anti-MAP antibiotic treatment is effective, but needs some larger-scale studies as well.

# 2.3.2. Autoimmunity can predispose to infectious diseases

During the autoimmune response, neutralization of important immune system components occurs by the autoantibodies, which are essential in mounting anti-microbial responses. These autoantibodies might either enhance current infectious disease(s) or susceptibility of the host to more infected microbes such as bacteria, viruses, and opportunistic fungus. For instance, cytokines – important regulatory molecules of the innate and adaptive immune system, plays an important role in increasing the anti-microbial response. In this context, some cytokines namely INF-c and IL-2 are produced by the Th1 cells and provide protection against intracellular virus and bacteria (for e.g. *Mycobacterium* *spp.* and *Salmonella sp.*). There are some classical examples of other cytokines also, which can provide protection against extracellular pathogens, *viz.*, Th-17 secrete IL-17, IL-21, and IL-22 that provided protection from *Candida* (a fungus) [(Aimanianda et al. 2009, Bettelli et al. 2008, Kolls & Khader 2011, Zhu & Paul 2008)](https://paperpile.com/c/a3wH5Q/5YUCi+eJxrB+LXY7D+c91Ou). Therefore, any hamper in cytokines activity would directly affect the defense mechanism, such as use of neutralizing antibodies against these cytokines affects the cellular functions and elimination of pathogens and predisposes the host to infectious diseases. Autoimmune patients with Crohn’s disease, rheumatoid arthritis or psoriasis (a skin disorder) treated with TNF-**𝛂** – monoclonal antibodies, increase vulnerability to mycobacterial, listerial, and viral infections [(Dinarello 2003, Maródi & Casanova 2010, Winthrop & Chiller 2009)](https://paperpile.com/c/a3wH5Q/4AKBT+SK9hy+l5vTJ).

# 2.3.3. Therapeutic options for autoimmunity-associated infectious diseases

To target autoimmunity-associated infectious (AAI) diseases most of the strategies should be intended to inhibit both infection as well as autoimmune response, such as blocking autoantibody-producing B cells and neutralizing autoantibodies are the most suitable therapeutic. The combination of immunosuppressive treatment and antimicrobial agents are the most appropriate therapeutic for AAI diseases. Alternatively, some strategies such as plasmapheresis i.e. removal of autoantibody or providing exogenous cytokines, to enhance autoantibody development, could also be used to treat AAI diseases, but the major disadvantage is that they are not eliminating the autoantibody source (i.e. autoantibody producing B-cells and plasma cells). B-cell targeted therapies can be used to eliminate antibodies producing B-cells, as they decrease the total immunoglobulin level and lead to serious infection predisposition, but do not kill the antibody producing plasma cells.

Furthermore, polyclonal intravenous immunoglobulin (IVIg) in combination with anti-microbial agents would be more efficient and safe therapeutics to target diverse AAI diseases [(Kazatchkine & Kaveri 2001, Tha-In et al. 2008)](https://paperpile.com/c/a3wH5Q/DoLtG+Ah49f). To maintain the immune tolerance and to suppress autoimmunity, IVIg targets both soluble mediators and cellular component of the autoimmunity, and there are several mechanisms through which it inhibit the diseases such as initiate B cell tolerance, regulation of immunoglobulin tolerance, neutralization of anti-cytokine autoantibodies by broad-spectrum anti-idiotypic antibodies, suppression of innate antigen presenting cells and inhibition of T cell help to B cells, and expansion of CD4+CD25+ regulatory T cells [(Kazatchkine & Kaveri 2001, Tha-In et al. 2008)](https://paperpile.com/c/a3wH5Q/Ah49f+DoLtG). However, an effective dose regime and duration of IVIg therapy determination is the major concern. Although, for several autoimmune and inflammatory diseases a combination of IVIg and B cell–targeted therapies are proved as successful therapy. Although, triple medication – combination of B-cell targeted therapies, antimicrobial agents, and IVIg would be considered most suitable therapies to target different AAI diseases [(Ahmed et al. 2006, Vo et al. 2008)](https://paperpile.com/c/a3wH5Q/miCpg+ftAOn).

# 2.4. *Mycobacterium tuberculosis* (Mtb): a case study

The process of molecular mimicry is also well known in*Mycobacterium**tuberculosis* (Mtb). Elkington et. al. (2016) have proposed that autoimmunity is a critical and overlooked process of TB pathology, and present clinical and experimental observations also support this hypothesis. For example in nearly half of TB patients’ autoantibodies responsible for wegener's granulomatosis and systemic lupus erythematosus have been observed [(Elkington et al. 2016)](https://paperpile.com/c/a3wH5Q/HZU7I).

Few other autoimmune diseases such as inflammatory bowel disease, behçet's disease, ankylosing spondylitis, crohn's disease, ulcerative colitis, and sarcoidosis have also been associated with the TB pathogenesis. Using differential gene expression analysis among patients with TB and patients with autoimmune or infectious diseases, Clayton et al. suggested that combination of infection and autoimmune disease signatures could explain 96.7% of the differentially expressed TB signature [(Clayton et al. 2017)](https://paperpile.com/c/a3wH5Q/EOvqr). On the basis of this observation they suggested that pathology in TB results from an interplay between infection and a currently unrecognized autoimmune process.

Though, it was observed that immunosuppressive therapies (mentioned in section 4.1.3) allow Mtb to proliferate [(Proal & Marshall 2018)](https://paperpile.com/c/a3wH5Q/2q89t). Similar observation (pathogenesis enhancement under immunosuppressive medications) extended to some other microbiome pathogens [(Finlay & McFadden 2006)](https://paperpile.com/c/a3wH5Q/DMXqx). It was reported that immunosuppressive therapies are majorly responsible for ‘microbial dysbiosis’ - microbial imbalance or maladaptation on or inside the body, and also associated with almost every autoimmune condition [(Diaz et al. 2013, Nellore & Fishman 2016)](https://paperpile.com/c/a3wH5Q/LsSVj+F9ki1). Some experimental evidence also suggests that patients treated with immunosuppressive therapeutic shows a high rate of relapse and co-morbidity.

In contrast, treatments which can target pathogens at the root of the disease process are needed to manipulate this complicated phenomenon. Herein, the mimicry inducing pathways can be used as potential targets, and drug repurposing would be a promising technique.

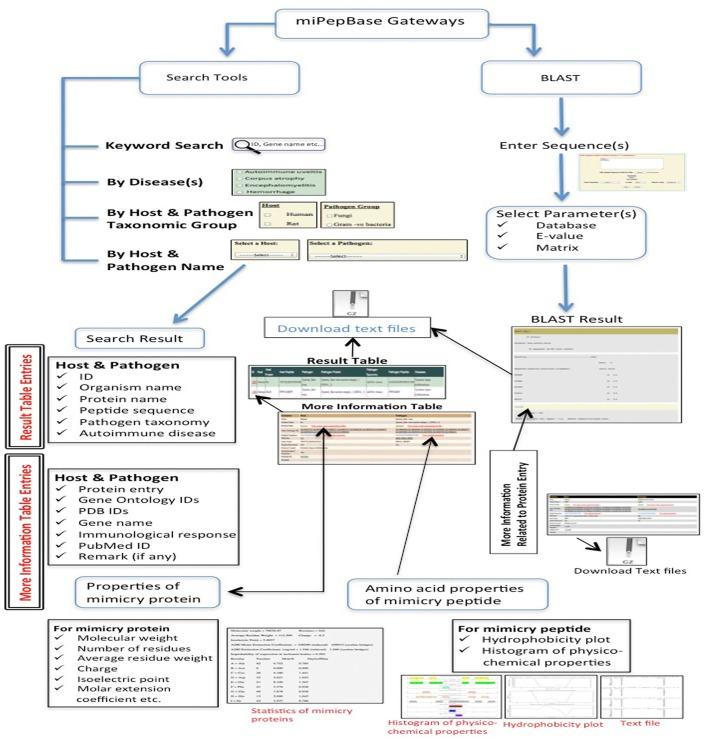
**Chapter 3**

# 

# 3.1. Introduction

Mimicry is a very common phenomenon in which a living being pretends to be what it is not. By adopting mimicry, an animal get protection by not hiding, rather being mistaken for something a predator will avoid because either it look dangerous or tastes bad. Hence, it is not surprising that similar strategy has been exploited at the molecular level as well. The obvious benefit molecular mimicry confers to pathogens is to fool the host's defenses and survive. The presence of a molecule in a pathogen that is similar with a host antigen could inhibit the immune response of the host against the pathogen because of the immune tolerance toward self-antigens [(Davies 1997, Gowthaman & Eswarakumar 2013)](https://paperpile.com/c/a3wH5Q/qwIIp+p4TgP). For example, *Helicobacter pylori* infection in human triggers two autoimmune diseases namely autoimmune gastritis and pernicious anemia. It occurs because activated CD4+ Th1 cells infiltrates into gastric mucosa and they cross-recognize the self-epitopes of H+K+ ATPase and *H. pylori* antigens [(D’Elios et al. 2004)](https://paperpile.com/c/a3wH5Q/slqTZ)There are number of well documented molecular mimicry events, using which bacteria, viruses, or parasites evade the host's immune response [(Oldstone 2005)](https://paperpile.com/c/a3wH5Q/PEA8y). The pathogen's protein having similar epitope to that of the host results in cross-reactivity that generates immunological response against self (i.e., host), which ultimately leads to autoimmune diseases Molecular mimicry and immune-mediated diseases [(Cusick et al. 2012, Oldstone 1998)](https://paperpile.com/c/a3wH5Q/Hfw6I+WvBKB)

The peptides, which display this property, are called mimicry peptides and the phenomenon is called molecular mimicry [(Davies 1997)](https://paperpile.com/c/a3wH5Q/qwIIp) The role of molecular mimicry in autoimmune disease was getting strengthen when it was observed that the antibody against the phosphoprotein of measles virus and Herpes simplex type I can cross-react with human intermediate filament protein vimentin [(Fujinami et al. 1983)](https://paperpile.com/c/a3wH5Q/LBBq9) Molecular mimicry can cause several immune-mediated disease such as Grave's disease [(Chen et al. 2001, Kohn et al. 2000)](https://paperpile.com/c/a3wH5Q/yDzXo+h9jhK), Insulin-dependent diabetes [(Hiemstra et al. 2001, Rose & Mackay 2000)](https://paperpile.com/c/a3wH5Q/AKFCg+vkFzp), Multiple sclerosis [(Banki et al. 1994, Rose & Mackay 2000, Talbot et al. 1996, Wucherpfennig & Strominger 1995)](https://paperpile.com/c/a3wH5Q/e9Bts+lu99O+6E11U+AKFCg). Peptic ulcer [(Appelmelk et al. 1996)](https://paperpile.com/c/a3wH5Q/MECL4), Rheumatoid arthritis [(Balandraud et al. 2004, Bridges 2004, Tiwana et al. 1999)](https://paperpile.com/c/a3wH5Q/jIiK5+wIK1H+CU8sp)), Systemic lupus erythematosus [(Kaufman et al. 2003, McClain et al. 2005, Rönnblom & Alm 2001)](https://paperpile.com/c/a3wH5Q/tbnpR+4LeCk+8vuSq), Myocarditis [(Ang et al. 2004, Gauntt et al. 1995, Huber et al. 1994, Neu et al. 1987, Schulze & Schultheiss 1995)](https://paperpile.com/c/a3wH5Q/QgsvU+H9sKV+1J254+Mw78O+MfCtP), and cancer as well, by modulating key signaling pathways, such as those involving Ras [(Guven-Maiorov et al. 2016)](https://paperpile.com/c/a3wH5Q/gffNr). A number of studies have deciphered various prospects and aspects of molecular mimicry, but these are scattered in numerous research papers. Compilation of the available information from literature can greatly facilitate the researchers who work in this domain. At present there is no data repository, which contains all the information related to autoimmune diseases caused due to molecular mimicry because piecing, together of this scattered data and discerning the accompanying details is complicated and tedious. To the best of our knowledge, only one database namely mimicDB [(Ludin et al. 2011)](https://paperpile.com/c/a3wH5Q/OSJm) is available which provides information about proteins or epitopes involved in host-pathogen interactions. But mimicDB is restricted to information pertaining to only a few human parasites. Also, the mimicry candidates of mimicDB were predicted through a computational pipeline.In the present study we have reported a freely accessible database, which can serve as a comprehensive and high quality resource of peptides involved in molecular mimicry. We have also incorporated the information related to autoimmune diseases as well as in-depth information about mimicry peptide and proteins. The database is named, **miPepBase (Mi**micry **Pep**tide Data**base**), which is available at<http://proteininformatics.org/mkumar/mipepbase>. All molecular mimicry based autoimmunity events compiled in miPepBase were experimentally verified by the respective researchers and are supported by peer-reviewed publications. **miPepBase** is an open access database that provides comprehensive information about the mimicry proteins and peptides of both host (and model) and pathogen. The information includes the names of host and pathogen proteins, sequences of mimicry peptide, autoimmune disease caused due to mimicry peptide, gene ontology information of the protein, PDB ID of the structure of protein (if present), type of immunological response generated by mimicry peptide and much more. We anticipate that miPepBase will help researchers to generate new hypothesis about different aspects of molecular mimicry and also act as a unified resource of information about molecular mimicry. The miPepBase can be searched using keyword(s) or by autoimmune disease(s) or by a combination of host and pathogen taxonomic groups or their names. The database also includes BLAST search tool to facilitate sequence similarity search against the mimicry proteins and/or peptide contained in it. Each miPepBase entry is also linked to many popular global repositories such as UniProt [(Apweiler et al. 2004)](https://paperpile.com/c/a3wH5Q/3tfSK), PDB [(Berman et al. 2000)](https://paperpile.com/c/a3wH5Q/d2fYk), EMBL-EBI QuickGO [(Binns et al. 2009)](https://paperpile.com/c/a3wH5Q/RuPb6), and PubMed. MiPepBase also provides information about physicochemical properties of proteins containing mimicry peptides, which might be helpful in predicting the nature of protein and optimization of its expression. The basic architecture of miPepBase is shown in Figure 3.1. The data of miPepBase can also be downloaded in text file. Overall, mimicry peptides which are compiled in miPepBase might help in opening new gateways to explore the role of molecular mimicry in autoimmune diseases that are yet unaddressed. It is anticipated that miPepBase would be helpful in understanding the details of molecular mimicry and expedite the process of disease detection, diagnosis, prognosis, and even deciding the therapeutic regimen of autoimmune disease.

**Figure 3.1.** Architecture of miPepBase. (Adapted from Garg, A., et al. (2017))

**3.2. Materials and method**

### **3.2.1. Data collection and compilation**

The main aim of miPepBase was to collect, compile and curate all the information related to autoimmune disease caused by molecular mimicry. Therefore, experimentally verified data was collected after an extensive search of published research papers with the help of PubMed and Google Scholar using keywords “molecular mimicry,” “host-pathogen cross-reactivity,” and “autoimmune diseases.” We also mined other additional relevant information such as gene and protein names, mimicry peptide sequence, name of autoimmune diseases, and immunological response by T-cells or antibodies. The information regarding proteins, taxonomic classification of pathogen, gene ontology information, PDB ID, annotation status of protein (review status) and protein sequences was obtained from the UniProt protein repository. The miPepBase also provides PubMed link with each entry from which the molecular mimicry and autoimmune disease information was extracted.

### **3.2.2. Web interface and database architecture**

The inner framework of miPepBase is built using MySQL ([http://www.mysql.org](http://www.mysql.org/)), Perl ([http://www.perl.org](http://www.perl.org/)), and Apache ([http://www.apache.org](http://www.apache.org/)) on Cent OS Linux platform. The interface component consists of webpages designed in HTML/CSS in a Linux environment. To provide convenience in usage, the database was developed in a user-friendly manner. The “Browse” and “Search” options were provided to search and access the information content of miPepBase. The home page of miPepBase has a very short introduction about molecular mimicry based autoimmune diseases. It also provides a brief description of the database content and clickable icons with direct links to the database and its different utilities.

### **3.2.3. Database accessibility**

The miPepBase provides interactive access to the data and the users can connect and access the database using any one among different search options. The search options have been designed in a simple and intuitive manner so that the users can search the database either by keyword or predefined combinations of fields (advanced search).

**Keyword search** assists users to search the database by following fields: database ID or organism's name or protein's name or entry or autoimmune disease or UniProt ID or taxonomic classification or gene ontology ID or PDB ID or peptide sequence or PubMed ID. It also permits free-floating Google like search over entire database.

**Advanced search** provides three different types of search options for users to access the data: First, search by one or multiple autoimmune disease(s) caused due to molecular mimicry. Second, search on the basis of host and pathogen taxonomic group, which allows users to explore one or multiple host(s) and pathogen taxonomic group(s) involved in molecular mimicry. The third and last option of advanced search is a drop down menu of host and pathogen name, which allows searching restricted to a specific set of host and pathogen. Irrespective of the mode of search chosen to query the miPepBase, the search result will be displayed in the tabular format. In the search result, the ID (shown in red color) is a clickable link and can display detailed information of corresponding entry. All the information can be downloaded in the text format, using the “download button” in result table. Additionally, different information related to protein sequence, structure, gene ontology and source of article, were linked to UniProt, RCSB PDB, EMBL-EBI QuickGO, and PubMed, respectively. A detailed step-by-step manual is also provided to assist users in smooth and efficient searching of miPepBase.

### **3.2.4. Tools integrated in miPepBase**

Different tools are also incorporated in the miPepBase to help users to search related proteins and/or peptides and analyze their different physicochemical properties. BLAST searches similar sequence(s) with in the database [(Altschul et al. 1997, 2005)](https://paperpile.com/c/a3wH5Q/97kJM+QiN6Y)while pepstats and pepinfo utilities of EMBOSS package provides information about physicochemical properties of protein and peptides [(Rice et al. 2000)](https://paperpile.com/c/a3wH5Q/1PgN6). The information derived from these tools might be helpful in predicting the nature of protein and optimization of protein expression.

**Pepstats** was used to calculate physicochemical properties of amino acids (such as molecular weight, number of residues) present in mimicry protein.

**Pepinfo** was used to calculate properties of mimicry peptide which include two types of plots: (i) Hydrophobicity plot (on the basis of Kyte and Doolittle parameters) and (ii) Histogram of presence of amino acid with the physico-chemical properties such as tiny, small, aliphatic, aromatic, non-polar, polar, charged, positive, and negative.

**Basic Local Alignment Search Tool (BLAST):** It is incorporated to find homologous sequence(s) and similar peptide(s) present within miPepBase database. User has to simply paste the sequence in the text box or upload sequence in the FASTA file to find similar sequence(s). Option to specify search parameters like database, *E*-value cutoff and alignment scoring matrix value is also present. The default cut-off *E*-value is 100 and alignment-scoring matrix is BLOSUM62. In the miPepBase BLAST tool, four different types of databases namely Host protein, Host peptide, Pathogen protein, and Pathogen peptide are present. Hence, similarity search can be carried out against any of the four databases.

# 3.3. Results

### **Data statistics and content**

In the miPepBase, only experimentally verified mimicry peptides from published papers are incorporated. The first release of miPepBase has 261 entries in total. It does not mean that miPepBase contains 261 host-pathogen peptide pairs. This is due to existence of multiple mimicry peptides in a single protein. Analysis of the miPepBase data shows that in both host and pathogen proteins more than one stretch of amino acids might be involved in molecular mimicry. The following information is associated with each entry:

* **ID:** It is a unique identifier assigned to each entry of the miPepBase database. Each ID is linked to the detailed information of that entry, which includes details of host and pathogen proteins, their gene ontology information, PDB ID of structure (if known), gene name, annotation status of protein (reviewed/not reviewed), PubMed ID, and remark (if any).
* **Organism's name:** With each event of molecular mimicry two different organisms are associated. Organism in which autoimmune response is generated was designated as host. Organism, which encodes the mimicry peptide, was designated as pathogen.
* **Protein names:** Two different proteins are associated with each event of molecular mimicry. One that is encoded by the host and second which is encoded by the pathogen. Names of both the proteins are present with each entry.
* **Peptide sequence:** This contains the stretch of amino acids (the peptide) present in both host’s and pathogen’s protein that actually leads to molecular mimicry.
* **Pathogen taxonomic group:** Organisms from all taxonomic groups such as bacteria, viruses, fungi, and protozoa exhibit molecular mimicry. MiPepBase contains information of molecular mimicry based autoimmunity events caused by organisms from all taxonomic groups.
  + Broadly, pathogens are divided into four taxonomic groups namely bacteria, fungi, protozoa, and viruses. Bacteria is further subcategorized into gram-positive, gram-negative, and others i.e., diderms. Further, viruses are categorized according to the classification system purposed by David Baltimore (reviewed in [(Baltimore 1971)](https://paperpile.com/c/a3wH5Q/W5j2o)).
  + Baltim, namely retro transcribing virus, dsDNA virus, dsRNA virus, and ssRNA virus. The total numbers of entries belonging to pathogens of different categories is shown in Figure 3.2.
* **Autoimmune disease:** This field provides the information about disease caused due to molecular mimicry. Our analysis revealed that very diverse types of autoimmune diseases might occur due to molecular mimicry. Data content of miPepBase shows total 23 types of autoimmune diseases are associated with molecular mimicry. Multiple sclerosis was the most frequent disease followed by encephalomyelitis. The different types of autoimmune diseases and the number of times they were associated with molecular mimicry is shown in Figure [​](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5660332/figure/F2/)3.2 (B).

# 

**Figure 3.2. Data statistics** (A) Based on pathogen taxonomic group, (B) Based on autoimmune disease. (Adapted from Garg, A., et al. (2017))

# 3.4. How to search query into miPepBase?

#### **3.4.1. Using a keyword**

Any data in miPepBase can be search and access by five different ways. It is illustrated here using one protein (UniProt accession number **P10809)**. Users can get the information associated to this protein by querying miPepBase submission of UniProt accession number as a keyword to the “Keyword search option” (Figure [​](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5660332/figure/F3/)3.3 (A)) and click the search button (Step a1). The search result page showed a single hit and the information related to P10809 protein was presented in tabulated form. The search result contains following information: unique miPepBase ID (1217), host name (human), host protein name (HSP60), host mimicry peptide sequence (HRKPLVIIAEDVDGE), pathogen name (*Mycobacterium bovis)*, pathogen protein name (HSP65), pathogen taxonomy (Gram positive bacteria), pathogen mimic sequence (AGKPLLIIAEDVEGE), and autoimmune disease (Rheumatoid arthritis) caused due to host and pathogen cross reactivity. All these details can also be downloaded as text file (Step a2). More detailed information related to P10809 can be retrieved through miPepBase ID of P10809 (i.e., 1,217, displayed in red font in the search table) (Step a3). Further, it will give more information about the host's and pathogen's: protein entry (host-P10809 and pathogen-P0A521), gene ontology (available for both), PDB ID (host- 4PJ1 and pathogen-NA), gene name (host-HSPDI, HSP60 and pathogen-groL2, groEL2, groEL2, hsp65, Mb0448), protein reviewed (host-yes and pathogen-yes), immunological response (Helper T cell), PubMed ID (1577070), and remark (NA). In addition to these details the miPepBase also provide direct link to UniProt, EMBL-EBI, RCSB PDB, and PubMed. All information described above can also be downloaded as “Text File” (Step a6).

Apart from above described information users can also get the amino acids composition profiles for P10809 (host's protein) and P0A521 (pathogen's protein) entries and their hydrophobicity graph and other physico-chemical information for mimicry peptides through “View amino acids composition profile” (Step a4) and “View peptide properties” (Step a5), respectively. All graphs and text file related to physico-chemical properties of protein and peptide can be downloaded in text format.

#### **3.4.2. By disease**

To retrieve the information related to mimicry proteins involved in a particular set of autoimmune diseases, users could use an advanced search option i.e., “Search by Diseases.” This option lists a set of disease caused due to molecular mimicry and whose information is present in miPepBase. Here, it is demonstrated using **Rheumatoid arthritis** as an example (Figure [​](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5660332/figure/F3/)3.3 (B)). On selection of rheumatoid arthritis as the disease whose information is desired (Step b1 and b2), search result page (Step b3) would be displayed. The search page would list the information related to proteins involved in the rheumatoid arthritis in a tabulated form. The information content and ways to navigate different sections remain same (Step a3–a6) as explained above for P10809 protein using Keyword search option.

#### **3.4.3. By host and pathogen taxonomic group**

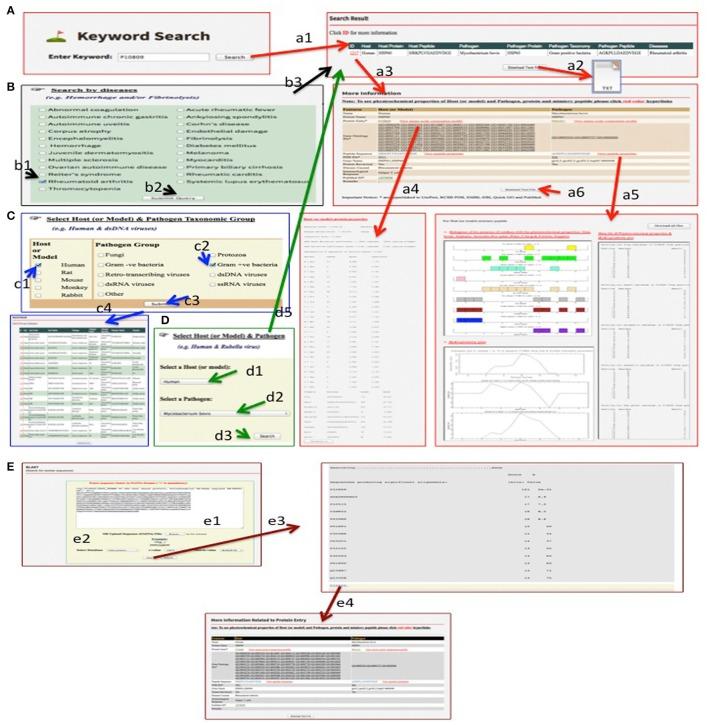
This option provides a list of pathogens and host taxonomic group within which the search will be restricted. This search option gives an easy way to do comparative analysis among mimics encoded by different pathogens of same or different taxonomic group(s) (Figure 3.3 (C)). Searching (Step c3) with “**Human**” **as host** (Step c1) **and “Gram-positive bacteria” as pathogen taxonomic group** (Step c2), total 19 entries related to gram-positive bacteria group (Step C4). Here also the presentation of search result data and further information (from Step a3 to a6) were remaining same as discussed for above two searching methods.

#### **3.4.4. By host and pathogen name**

The information related to event of cross reactivity between a specific host and pathogen that leads to autoimmune disease(s) can be achieved using another advanced options i.e., “Select host and pathogen.” The names of host and pathogen can be selected from the dropdown menu present in this section. Here, it is exemplified (Figure 3.3 (D)) using **Human as host** (Step d1) **and** ***Mycobacterium bovis* as pathogen** (Step d2). After submission of query (Step d3) a result page would be display that contains the search result information in tabulated form. The information content of search page will remain same as explained earlier for keyword search (Step a3–a6).

#### **3.4.5. By blast search**

This is not a direct way to search the data content of miPepBase. Rather it searches similar sequences and peptides in the miPepBase. The BLAST search option is available at menu bar (Figure 3.3 (E)). The query sequence in FASTA format can either be pasted in the text box or uploaded as sequence file (Step e1). Three parameters have to be optimized for efficient BLAST search (i) the database in which related sequence will be searched; (ii) *E*-value, and (iii) Scoring Matrix. The default *e*-value and scoring matrix are 100 and BLOSUM 62, respectively (Step e2). As shown in section E, whenP10809 protein sequence was searched against host protein database, total 13 hits were obtained which are arranged on the basis of ascending *e*-value (Step e3). Also every BLAST hit protein entry is further linked (in blue color) to detailed information page, which provide tabulated detailed information of corresponding BLAST hit (Step e4). These information are the same as described above for keyword search option (Step a4–a6).



**Figure 3.3. Process of stepwise data retrieval and analysis in miPepBase.** The user can search query with following options: (A) Keyword search, (B) Search by disease, (C) Search by host and pathogen taxonomic group, (D) Search by host and pathogen name. The search from (A–D) options display search result table and from that user can select the entry/displayed result for further detailed analysis. Sequence based search can also be searched by (E) BLAST search option and each hit is further linked to its details information page. The detail of result obtained from search options (A–E) is displayed by corresponding small case (the number indicates step number). From example a1–a6 denotes the results that can be obtained using keyword search option (A). (Adapted from Garg, A., et al. (2017))

**3.5. Discussion**

During the last few years, much active research and experimental verification has shed light on various aspects of molecular mimicry and it's role in autoimmune diseases. With the passage of time, number of autoimmune diseases caused due to molecular mimicry is increasing. Since, a unified repository of the available information related to molecular mimicry based autoimmune diseases is not available, hence we have built a database (miPepBase) which not only contains the information regarding proteins and peptides associated with the process, but several other important details also. In-depth analysis of this information might lead to the elucidation of mechanisms of autoimmune diseases controlled by mimicry peptides. Each entry in the miPepBase database is linked to many other molecular biology data repositories. Further, the database also includes inbuilt tools, which can help to fetch other relevant information related to the mimicry proteins and peptides. As more data will accumulate by the use of high throughput molecular, genomic and metagenomic methods, we anticipate that the release of miPepBase will facilitate comprehensive analyses of different factors involved in autoimmune diseases caused by the mimicry peptides. We also hope that miPepBase would be helpful for the scientific community in understanding the host-pathogen interactions, as well as how the pathogens evade host immune systems.

## 3.6. Comparison with other available database of antigenic peptides

Several web-based antigen/epitope databases are available the content of which is freely available to the users. A brief description of **MimicDB** along with comparison with miPepBase is as follows:

### **MimicDB**

mimicDB [(Ludin et al. 2011)](https://paperpile.com/c/a3wH5Q/OSJm) is a database of linear amino acid epitopes derived from a comparative genomics approach. These epitopes were predicted to be a potential molecular mimicry peptide and derived from a computational prediction pipeline. Further mimicDB is focused on a few selected human endoparasites namely *Brugia malayi, Schistosoma mansoni, Plasmodium falciparum, Leshmania major, Cryptosporidium parvum, Trichomonas vaginalis*, and *Trypanosoma cruzi*. In miPepBase the information is not restricted to any particular class of pathogen and/or disease. It contains information related to all autoimmune diseases caused by pathogens, which may belong to viruses, or prokaryotes, or eukaryotes. miPepBase host's and pathogen's mimicry peptides were curated from literature. The respective researchers have already experimentally established the role of these mimicry epitopes in generating autoimmune disease.

## 3.7. Limitations and future prospects

Although, we have made outmost effort to compile all available data at one place, it cannot be claimed that miPepBase contains information about each and every peptide/protein involved in molecular mimicry based autoimmune diseases. It is certainly possible that few peptides might have been missed and not included in the miPepBase. In future, we would make our best efforts to include the missing as well as newly added data in miPepBase. The motivation behind establishment of miPepBase was to establish a knowledgebase for proteins/peptides involved in molecular mimicry. We will continue to add new information, which may include but not limited to interaction partners of mimicry proteins and their role in disease. This will enable us to provide a platform for study of the mimicry peptides and pathways through which they trigger autoimmune diseases. We believe the miPepBase database would helpful to the scientific community in exploring the various prospect and aspects of molecular mimicry.

**3.8. Database update**

An important aspect of any database is to keep it up to date by adding new data. We would constantly add information about newly discovered peptides, which exhibit molecular mimicry and cause autoimmune diseases.

## 3.9. Accessibility and data download

The database and its contents are freely accessible without any restriction at<http://proteininformatics.org/mkumar/mipepbase>.

The results described in this chapter have been published as - Garg et al. miPepBase: A database of experimentally verified peptides involved in molecular mimicry. ***Frontiers in Microbiology***, 8, 2053, 2017

**Chapter 4**

# 4.1. Introduction

Molecular mimicry can be defined as similarities between the sequences/structures of host and microbial proteins/peptides. These similarities might cause cross-activation of the host autoreactive T- or B-cells against self-epitopes, resulting in tissue and/or organ destruction and, ultimately autoimmune disease(s). Although host genetics play an important role in induction of host autoimmune response, several epidemiological and molecular evidences suggest that bacteria and viruses might be the principal environmental triggers of autoimmunity [[1–5]](https://paperpile.com/c/QlJf8f/hz0Vq+MFhwA+gYLN7+RTlcz+vXf0u). Several studies have associated different bacteria and viruses with autoimmune diseases. Some common autoimmune diseases triggered by bacteria and viruses include multiple sclerosis [[6]](https://paperpile.com/c/QlJf8f/t375), type 1 diabetes mellitus [[7]](https://paperpile.com/c/QlJf8f/YZeAj), autoimmune uveitis [[8]](https://paperpile.com/c/QlJf8f/9ghwT), encephalomyelitis [[9]](https://paperpile.com/c/QlJf8f/7Fuqv), inflammatory bowel disease [[10,11]](https://paperpile.com/c/QlJf8f/MwyJE+mPeG8), Crohn’s disease [[12,13]](https://paperpile.com/c/QlJf8f/JPDjM+Kjxrk), sarcoidosis etc.

Though researchers have studied various aspects of molecular mimicry and a few studies have reported some structural characteristics of viral mimicry proteins [[14, 15]](https://paperpile.com/c/QlJf8f/DakD+WsJ3) to the best of our knowledge, structural and functional characteristics of mimicry proteins of bacteria, viruses and host have not been studied in detail. Thus, in the present study we have performed a detailed *in silico* characterization of structural and functional characteristics of bacterial, viral and host mimicry proteins. Initially, the structural features of the mimicry proteins were examined by investigating if the host and pathogen mimicry proteins were ordered or disordered. Next, we examined the intrinsically disordered regions (IDRs) of the mimicry proteins because even though IDRs do not have a well-defined three-dimensional structure, they facilitate key cellular functions [[16–20]](https://paperpile.com/c/QlJf8f/Z5lJ6+Vx04A+CHGl5+8S0ME+UoRwT). The importance of IDRs can also be understood from the fact that pathogenesis of viruses has been strongly correlated with the presence of IDRs in their proteins [[21]](https://paperpile.com/c/QlJf8f/bOp6l). Within IDRs some conserved elements are present like, molecular recognition features (MoRFs) and short linear motifs (SLiMs). MoRFs are amino acid stretches of 10-70 amino acids which can undergo a disorder-to-order transition upon binding to their partners and are implicated in protein-protein interactions [[22]](https://paperpile.com/c/QlJf8f/zeeVa). SLiMs are short stretches of amino acids (3-10 amino acids) which are functionally diverse and mediate signalling interactions [[23,24]](https://paperpile.com/c/QlJf8f/syMId+8O37J). Several studies have indicated that viruses manipulate the host cellular machinery by mimicking SLiMs and MoRFs of the host proteins [[14, 15]](https://paperpile.com/c/QlJf8f/DakD+WsJ3). Similarly, several pathogenic bacteria exhibit molecular mimicry with the host SLiMs to propagate and sustain themselves inside the host [[25]](https://paperpile.com/c/QlJf8f/2Sqoh). Besides the presence of MoRFs and SLiMs, disordered regions are often characterised by low complexity regions (LCRs) made up of homo-polymeric repeats of a single amino acid or hetero-polymeric short repeats of a few amino acids residues [[26]](https://paperpile.com/c/QlJf8f/inicT).  Initially LCRs were considered to be primarily disordered but a few recent studies suggest that some LCRs can have a regular secondary structure, also [[27]](https://paperpile.com/c/QlJf8f/rubv). In the proteins, LCRs are associated with several important functions like, antigen processing and diversification, protein-protein interactions etc [[28]](https://paperpile.com/c/QlJf8f/S87lN). Collectively, SLiMs, MoRFs and LCRs help proteins in adopting several dynamic functional structures which enables their interaction with multiple binding partners [[29]](https://paperpile.com/c/QlJf8f/XCsg). In the present study, we have examined the presence and characteristics of disordered regions like MoRFs, SLiMs and LCRs in the bacterial, viral and host mimicry proteins. The functional annotation of the mimicry proteins was performed using the Gene Ontology (GO) annotations retrieved from the Gene Ontology Consortium [[30]](https://paperpile.com/c/QlJf8f/jSlrU). Additionally, the probable implications of the structural and functional characteristics of the mimicry proteins/peptides are also discussed.

# 4.2. Material and Methods

## 4.2.1. Retrieval of experimentally validated mimicry proteins from miPepBase

In the present study, the information on bacterial and viral mimicry proteins along with the host mimicry proteins was retrieved from an earlier developed database of experimentally verified mimicry proteins, miPepBase [[31]](https://paperpile.com/c/QlJf8f/LuW3e). The mimicry epitopes of the host and pathogen proteins were named as host-mimitope and path-mimitope, respectively and the proteins were named as host-protein and path-protein, respectively.

## 4.2.2. Structural characterization of host and pathogen mimicry proteins

**4.2.2.1.** ***Order/disorder propensity of amino acids in the host- and path-proteins***

The order/disorder predisposition of amino acids in the host- and path-proteins was predicted using the consensus in three different disorder predictors namely, DISOPRED (version 3.16), IUPred (version 1.0) and PONDER VSL2. DISOPRED is a hybrid predictor based on SVM, neural network and nearest neighbour classifiers [[32]](https://paperpile.com/c/QlJf8f/R2Ovp). For each amino acid, DISOPRED gives a score between 0-1, with 0.50 as the threshold boundary. Amino acids with a DISOPRED score of ≤0.50 are considered as ordered while with a score >0.50 as disordered. IUPred predicts the disordered/unstructured regions in a protein sequence based on total pairwise inter-residue interaction energy and on the assumption that intrinsically unstructured protein sequences do not fold due to their inability to form a sufficient number of stabilizing inter-residue interactions [[33]](https://paperpile.com/c/QlJf8f/GW7BD). Similar to DISOPRED, the score of IUPred prediction also ranges from 0 (complete order) to 1 (complete disorder) with a score above 0.5 indicating disorder. PONDR® FIT (Predictor of Natural Disordered Regions) includes six different types of predictors. We used PONDR® VSL2 since it is considered as the most accurate [[34]](https://paperpile.com/c/QlJf8f/Wv9vY). In PONDR, any residue with a score ≥ 0.5 is considered as disordered. In the present study, all the three IDPR predictors were used at default parameters. To address the variability in the predictions of the three predictors, the final disorder predisposition of each amino acid was calculated based on consensus in all the three predictions. An amino acid was annotated ordered/disordered based on consensus in the predictions of at least two of the three predictors. On the basis of consensus prediction, we calculated the percentage of disordered residues (PDR) in a protein by dividing the number of residues in a protein that were predicted as disordered by the total number of residues in that protein. On the basis of PDR, all the host- and path-proteins were divided into three categories: highly ordered (PDR < 10%), moderately disordered (10% ≤ PDR < 30%), and highly disordered (PDR ≥ 30%). Previously, several researchers have also used PDR to classify proteins as ordered/disordered [[35–38]](https://paperpile.com/c/QlJf8f/qcdS1+idO4J+CyX8j+ZUoDv).

## 4.2.2.2. Investigating the MoRFs, SLiMs and LCRs

The presence of MoRFs in the mimicry proteins and mimitopes were investigated using the MoRFchibi SYSTEM which contains three different modes of MoRF predictions, MoRFCHiBi, MoRFCHiBi\_Light, and MoRFCHiBi\_Web [[39]](https://paperpile.com/c/QlJf8f/TlnSM). In the present study, MoRFCHiBi\_Web mode was used, which though slower than the other two modes, makes highly accurate predictions [[39]](https://paperpile.com/c/QlJf8f/TlnSM). To predict SLiMs in the mimicry proteins and mimitopes ANCHOR [[40]](https://paperpile.com/c/QlJf8f/WGX5m) was used and to find LCRs, SEG [[26]](https://paperpile.com/c/QlJf8f/inicT) was used. For annotating LCRs in a protein sequence, SEG uses three numeric parameters *viz.* window length (L), trigger complexity (K1), and extension complexity (K2). The whole process of LCR identification by SEG completes in two-steps. First, it identifies a low complexity segment using a sliding window of length L amino acids that have local sequence complexity K1. Then all overlapping subsequences with sequence complexity K1 are merged in both directions till the complexity of a contig built by overlapping subsequences does not exceed K2. In the present work, MoRFCHiBi\_Web, ANCHOR and SEG were used at default parameters. A MoRF/SLiM/LCR was considered to overlap with the host or pathogen mimicry mimitope if at least half of the amino acid residues of mimitopes overlapped with these regions.

## 4.2.3. Functional enrichment analysis

We performed the GO term based functional enrichment analysis of host- and pathogen-proteins using GO mapping tools, called OWLTools’ Map2Slim (M2S) (https://github.com/owlcollab/owltools). The functional enrichment analysis was performed at all three levels of GO annotations namely, Molecular Function, Cellular Component and Biological Process.

# 4.3. Results

## 4.3.1. Benchmarking dataset

A total of 152 bacterial mimicry proteins and their corresponding 28 host proteins and 34 viral mimicry proteins and their corresponding 22 host mimicry proteins were retrieved from the miPepBase. The bacterial and their corresponding host mimicry proteins were named as Bacterial-set proteins and, viruses and their corresponding host mimicry proteins were named as Viral-set proteins. The Bacterial-set proteins were involved in 16, while the Viral-set proteins were involved in 12 different types of autoimmune diseases (Table E4.1).

## 4.3.2. Structural order/disorderliness in the host and pathogen mimicry proteins

To assess the extent of structural disorderliness, we categorized the mimicry proteins in three different categories based on the PDR. Our results revealed that 110 of the 152 bacterial (~73%) and 16 of the 34 (47%) viral mimicry proteins were ordered (PDR <10%). This implies that 27% of the bacterial mimicry proteins were disordered, while 53% of the viral mimicry proteins were disordered. Approximately, 75% of the host mimicry proteins of the Bacterial-set and ~72% of host mimicry proteins in the Viral-set showed moderate to high disorderliness (Table 4.1, E4.2). In the Bacterial-set, 50% of the host mimitopes and 12.5% of the bacterial mimitopes exhibited disordered regions. In the Viral-set, 36.3% of the host mimitopes and 26.4% of the viral mimitopes exhibited disordered regions (Table 4.1, E4.2).

## 4.3.3. MoRFs, SLiMs and LCRs in mimicry protein/peptides

In the Bacterial-set, 118 of the 152 bacterial and 26 of the 28 host proteins contained MoRFs, while in the Viral-set 33 of the 34 viral and 15 of the 22 host proteins contained MoRFs (Table 4.2, E4.2). Further, nine bacterial and five host mimitopes in the Bacterial-set contained the MoRFs. In the Viral-set, only three viral and four host mimitopes possessed MoRFs (Table 4.2, E4.2).

Analysis of the SLiM regions in the host and pathogen mimicry proteins revealed that 18 host and 43 bacterial proteins harbored SLiMs in the Bacterial-set. In the Viral-set, 11 host and 21 viral proteins harbored SLiMs (Table 4.2, E4.2). Further, nine bacterial and host mimitopes each in the Bacterial-set harboured SLiMs. In the Viral-set, four viral and five host mimtopes harboured the SLiMs (Table 4.2, E4.2).

In the Bacterial-set, 91 of the  bacterial proteins and 20 host proteins harbored LCRs, while in the Viral-set 30 viral proteins and 16 host proteins harboured LCRs (Table 4.2, E4.2). In the Bacterial-set nine bacterial mimitopes and two host mimitopes contained the LCRs, while in the Viral-set six viral mimitopes and one host mimitope contained the LCRs (Table 4.2, E4.2).

**Table 4.1.** Categorization of mimicry proteins of bacteria, viruses and host on the basis of PDR (percentage of disordered residues in protein)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | **Categorization of mimicry proteins on the basis of PDR** | | |
| **Mimicry proteins** | | **No. of ordered mimicry proteins (PDR <10%)** | | **Number of moderately disordered mimicry proteins (10% ≤ PDR < 30%)** | **No. of disordered mimicry proteins (PDR ≥ 30%)** |
| Bacterial-set proteins | Bacterial proteins | 110 | | 31 | 11 |
| Host proteins | 6 | | 8 | 14 |
| Viral-set proteins | Viral proteins | 16 | | 8 | 10 |
| Host proteins | 6 | | 8 | 8 |

**Table 4.2.** Distribution of MoRFs, SLiMs and LCRs in the mimicry proteins and mimitopes of bacteria, viruses and host

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Structural feature** | **Categorization of mimicry proteins** | | **Number of mimicry proteins** | **Number of mimitopes** |
| Molecular recognition features (MoRFs) | Bacterial-set proteins | Bacterial proteins | 118 | 9 |
| Host proteins | 26 | 5 |
| Viral-set proteins | Viral proteins | 33 | 3 |
| Host proteins | 15 | 4 |
| Short linear motifs (SLiMs) | Bacterial-set proteins | Bacterial proteins | 43 | 9 |
| Host proteins | 18 | 9 |
| Viral-set  proteins | Viral proteins | 21 | 4 |
| Host proteins | 11 | 5 |
| Low complexity regions (LCRs) | Bacterial-set proteins | Bacterial proteins | 91 | 9 |
| Host proteins | 20 | 2 |
| Viral-set | Viral proteins | 30 | 6 |
| Host proteins | 16 | 1 |

## 4.3.4. Functional characterization of the host and pathogen mimicry proteins

The GO annotations revealed that the host mimicry proteins of the Bacterial-set were involved in various biological processes like, anatomical structure development (11 proteins), ion binding (10 proteins), response to stress (7 proteins), immune system process (6 proteins), signal transduction (6 proteins), cell differentiation (6 proteins), cellular protein modification (6 proteins) and protein transport (5 proteins). Overall, the host proteins of Viral-set were also involved in similar functions like ion binding (9 proteins), anatomical structure development (8 proteins), biosynthetic process (8 proteins), immune system process (7 proteins), signal transduction (7 proteins), cell-cell signaling (7 proteins) and catabolic process (7 proteins) (Figure 4.1(a), Table E4.3).

Similar to the biological processes, in the molecular function category also, the overall function of bacterial and viral host proteins were similar, except that in the host proteins of Viral-set, a few functions like, transmembrane transporter activity were absent (Figure 4.1(b), Table E4.3). The bacterial proteins were involved in ion binding (49 proteins), biosynthetic process (32 proteins), cellular nitrogen compound metabolic process (23 proteins), carbohydrate metabolism (16 proteins), transmembrane transport (13 proteins), DNA binding (13 proteins) and DNA metabolic process (12 proteins). Functional enrichment of the viral proteins suggested that most of the proteins were a part of symbiont process (28 proteins), cellular nitrogen compound metabolic process (11 proteins), biosynthetic process (10 proteins), DNA binding (10 proteins), ion binding (8 proteins), immune system process (7 proteins) and membrane organization (7 proteins) (Figure 4.1(b), Table E4.3).

In the category cellular components, the host mimicry proteins of the Bacterial-set were  present in plasma membrane (10 proteins), cell (10 proteins) and protein-containing complex (10 proteins) while the bacterial mimicry proteins showed a significant presence in cytoplasm (32 proteins) followed by plasma membrane (15 proteins) and protein-containing complex (11 proteins). The functional analysis of Viral-set proteins revealed that host mimicry proteins were localized at the plasma membrane (13 proteins), cell (10 proteins), and cytoplasm (7 proteins) while no specific sub-cellular enrichment was observed for viral mimicry proteins (Figure 4.1(c), Table E4.3).

We also observed that the majority of the host mimicry proteins of the Bacterial- and Viral-set were multifunctional (Figure 4.1(a-c), Table E4.3). For example, molecular mimicry between the human Zinc transporter 8 protein (Uniprot Id: Q8IWU4) and MAP\_3865c protein of *Mycobacterium avium* subsp. *paratuberculosis* results in Type 1 diabetes mellitus. GO analyses of the host mimicry protein Q8IWU4 revealed its involvement in several biological processes like ion binding, stress response, immune system process, protein transport, homeostatic process, transmembrane transport, vesicle-mediated transport and cell-cell signalling. This might be a probable reason underlying its localization at different organelles like, plasma membrane, golgi apparatus and cytoplasmic vesicles (Table E4.3).

# Macintosh HD:Users:anjali:Desktop:Str_function:Final-draft-manuscript:Figure 1(a).jpg

**Figure 4.1.** Gene Ontology based functional annotation of bacterial, viral and host mimicry proteins. (a) Biological Process, (b) Molecular Function and (c) Cellular Component

# Macintosh HD:Users:anjali:Desktop:Str_function:Final-draft-manuscript:Figure 1(b).jpg

# Macintosh HD:Users:anjali:Desktop:Str_function:Final-draft-manuscript:Figure 1(c).jpg

# 4.4. Discussion

The current knowledge about the host and pathogen proteins involved in molecular mimicry is fragmentary [[25, 41]](https://paperpile.com/c/QlJf8f/9qvah+2Sqoh). Therefore, a detailed characterization of the host and pathogen mimicry epitopes and proteins can advance our understanding of the molecular mechanisms underlying this process. Thus, in the present study, we have performed a detailed *in silico* structural and functional analysis of bacterial, viral, and host mimicry proteins and mimitopes to understand some unexplored aspects of this process.

Though, the overall fraction of disordered residues varies widely in both bacterial and viral proteomes, it has been observed that in the majority of interactions between virus and host proteins, at least one of the two interacting protein partners is structurally disordered [[35]](https://paperpile.com/c/QlJf8f/qcdS1). In viruses, structure disorderliness might range from as low as 7.3% in the human coronavirus NL63 to as high as 77.3% in the avian carcinoma virus, while in bacteria it usually varies in a small range of 18-35% [[42]](https://paperpile.com/c/QlJf8f/PKGa1). A similar pattern of structure disorderliness was observed in our study, too. The number of viral mimicry proteins (53%) and mimitopes (27%) which exhibited disorderliness was much greater than bacterial mimicry proteins (27%) and mimitopes (12.5%). This observation can be explained by two plausible reasons. The high content of disordered proteins in viruses might enable them to carry out multiple functions [[43]](https://paperpile.com/c/QlJf8f/0nRwI). Also, a lack of well-defined structure and unrestrained conformations of viral proteins might facilitate adaptation inside the host.

With regards to the host, almost half of the host mimitopes involved in bacterial mimicry and only one-third of the host mimitopes in viral mimicry were disordered. Several studies have indicated that human proteins involved in the antiviral innate immune response [[43]](https://paperpile.com/c/QlJf8f/0nRwI), apoptosis, autophagy and programmed necrosis (necroptosis) possess an intrinsic disorder [[28]](https://paperpile.com/c/QlJf8f/S87lN). Interestingly, intrinsically disordered regions of the proteins are highly flexible and thus help the host cells in combating the viral invaders. Thus, it can be anticipated that most of the viruses and bacteria might preferentially select ordered regions of the human proteins for molecular mimicry because ordered regions of the proteins possess less structural flexibility and might be less involved in host immune responses. Moreover, this might be a probable underlying reason that only a small fraction of the mimitopes contained the disordered regions like SLiMs, MoRFs and LCRs.

Several studies have indicated that viruses hijack the host cellular machinery for integration, establishment and propagation inside the host [[44–46]](https://paperpile.com/c/QlJf8f/nKuER+IxMp5+QvOjN). Recent studies suggest that bacteria also alter the host machinery and use it for their own survival and propagation [[25, 47]](https://paperpile.com/c/QlJf8f/2Sqoh+c2d4W). It is well-known that metal-sequestering host-defense proteins and microbial metal acquisition machinery are important players in bacterial pathology and disease outcome [[47]](https://paperpile.com/c/QlJf8f/c2d4W). The host immune system counteracts the bacterial infections by reducing metal availability, but the pathogen outmaneuvers this by hijacking the host metalloproteins. In the present study too, we observed that the bacteria and viruses exhibited molecular mimicry with those proteins of the host which were involved in ion binding and signalling pathways.

To summarize, our results indicated that the most of the mimicry proteins and mimtopes of the bacteria, viruses and the host were ordered and, only a few mimitopes harboured the disordered regions like MoRFs, SliMs and LCRs. Also, the fact that most of the host mimitopes were ordered suggests that viruses and bacteria might preferentially select those regions of the host proteins for molecular mimicry which are ordered, and hence possess less structural flexibility and lesser immunological activity. Functional analyses indicated that both bacterial and viral mimicry proteins were involved in similar functions while the host mimicry proteins were multifunctional and mainly involved in ion binding, symbiont processes and signalling pathways. To the best of our knowledge, this is the first detailed *in silico* study on structural and functional aspects of bacterial, viral and host mimicry proteins and their plausible implications. We hope our report might serve as a useful platform for further studies on molecular mimicry. 

**Chapter 5**

# 5.1. Introduction

When macromolecules found on pathogens and in host tissues share structural, functional or immunological similarities, it is called molecular mimicry [(John Wiley & Sons, Ltd 2001)](https://paperpile.com/c/a3wH5Q/s4tKM). Molecular mimicry might occur in the form of complete identity or homology at the protein level or as similarity in sequences of amino acids and structure. Sequences based molecular mimicry plays an important role in immune response to infection and in autoimmune diseases. In order to attribute an autoimmune disease with molecular mimicry, certain criteria should be met: (1) there should be similarity between an epitope of host, microorganism or environmental agent, (2) antibodies or T cells cross-reactive with both epitopes must be detected in patients with the autoimmune disease, (3) there should be evidence of an epidemiological link between exposure to a microbe or an environmental agent and development of autoimmune disease and (4) autoimmune disease should be able to develop in an animal model when sensitized with the epitopes or, exposed to the environmental agent or infected with the microbe [(Peterson & Fujinami 2007)](https://paperpile.com/c/a3wH5Q/3gtqM). Many pathogens exhibit molecular mimicry with the host’s proteins and cause autoimmune diseases, thereof. These pathogens have been enlisted in Table E5.1. A detailed and explicit study on the role of molecular mimicry in microbial pathogenesis has not been conducted for most of the pathogens, except for a few fragmentary studies. For example, it was reported that group A *Streptococcus* andgroup B *Neisseria meningitides* use molecular mimicry to prevent induction of pathogen-specific immune response [(Nisini 2016)](https://paperpile.com/c/a3wH5Q/vzgjk). Autoantibodies responsible for wegener's granulomatosis and systemic lupus erythematosus have been observed in nearly half of the patients suffering from tuberculosis (TB) [(Kakumanu et al. 2008)](https://paperpile.com/c/a3wH5Q/KrpDm) A few other autoimmune disease such as inflammatory bowel disease, behçet's disease, ankylosing spondylitis, crohn's disease, ulcerative colitis, and sarcoidosis have been associated with pathogenesis of *M. tuberculosis* (Mtb) [(Elkington et al. 2016)](https://paperpile.com/c/a3wH5Q/HZU7I). Analysis of differential gene expression among TB patients and patients with autoimmune or infectious diseases indicated that a combination of infection and autoimmune disease signatures could explain 96.7% of the differentially expressed TB signatures [(Clayton et al. 2017)](https://paperpile.com/c/a3wH5Q/EOvqr). Autoimmunity has been considered a critical and overlooked process in pathology of tuberculosis (TB), with fragmentary study associated on the belief that TB an “infection-initiated autoimmune disease” [(Elkington et al. 2016)](https://paperpile.com/c/a3wH5Q/HZU7I). Autoimmunity has not been considered as a critical process in pathology of tuberculosis (TB). It continued to be an overlooked event with fragmentary studies.

Blocking the metabolic chokepoint has been used as a successful strategy for identifying new drug-targets against a particular organism [(Gupta et al. 2017, Rahman & Schomburg 2006)](https://paperpile.com/c/a3wH5Q/oCuTs+qIRqg). In the present review, we have described how blocking the chokepoint involved in production of pathogen’s mimicry proteins and their interaction partners can be used for discovery of novel targets against pathogens. In this review this approach has been explained using *M. tuberculosis* as the model organism. The initial step in this process involves identification of interaction partners of pathogen’s proteins (IPPP) involved in molecular mimicry with the host proteins. The homologs of host protein which might be present in IPPPs are removed, and chokepoints of the metabolic pathway are identified. Finally, drug candidates targeting the chokepoint proteins are selected from the DrugBank database and their efficiency and suitability is assessed. The steps involved in the process are shown in Figure 5.1.



**Figure 5.1.** The scheme of drug repurposing proposed against *Mtb*.In the figurewe have explained the complete process which is clustered into three major sections: (1) ***Interactome analysis*** includes protein data retrieval; collection of interacting proteins and removal of path-proteins which are homologous to human protein(s). (2) ***Filter potential target(s*)** that include mapping of mycobacterial nHIPPP dataset into their metabolic pathway(s) and search of possible chokepoint protein(s). If chokepoint proteins could able to pass through filters namely part of core proteome (A) or essential proteins (B). All chokepoint protein that crosses either filter is moved to the third step. (3) ***During drug repurposing*** chokepoints proteins were searched for effective ligand(s) and their interaction was analyzed after docking. In the last step *Mtb* homolog was searched for each chokepoint protein (C). List of databases & servers used during the whole process is 1: miPepBase, 2: STRING, 3: KEGG, 4: UniProtKB, 5: DEG, 6: DrugBank, 7: PatchDock, 8: LigPlot+ v.1.4 (Adapted from: Garg *et al* (2018))

**5.2. Material and method**

## 5.2.1. Data extraction

The experimentally verified events in autoimmune diseases caused due to molecular mimicry were obtained from a database developed by us earlier, miPepBase [(Garg et al. 2017)](https://paperpile.com/c/a3wH5Q/ElOF5). In brief, miPepBase is an indigenously developed, manually curated database containing information about proteins/peptides, which exhibit molecular mimicry and, autoimmune diseases, thereof. In the present work, epitopes of the pathogen and host proteins involved in molecular mimicry are referred as path-memitope and host-memitope respectively. Similarly, proteins carrying path-memitope and host-memitope are referred as path-protein and host-protein, respectively.

## 5.2.2. Protein-protein interaction search

The IPPP were found using the database, STRING [(Szklarczyk et al. 2017)](https://paperpile.com/c/a3wH5Q/5Kkbq). STRING contains information about protein interactions, established by experimental studies and by genomic analysis like domain fusion, phylogenetic profiling and gene neighborhood. We included only those interactions which scored ≥0.4.

## 5.2.3. Pathway mapping and determination of chokepoints in mycobacterial metabolic pathway

The 143 nHIPPP belonged to *M. leprae*, *M. avium* subsp. *paratuberculosis* and *M. tuberculosis.* Each nHIPPP was mapped in their corresponding metabolic networks in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [(Kanehisa 2000)](https://paperpile.com/c/a3wH5Q/nrKEf). KEGG is a database resource that cross-integrates genomic, chemical, and systemic functional information of an organism. Due to this, KEGG is widely used as a reference knowledge base for integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies.

## 5.2.4. Authentication of chokepoint targets and druggability of selected targets

The validation of essentiality of chokepoint proteins in mycobacterialmetabolic pathway was done by two ways: (a) homologs of chokepoint proteins were searched in all known mycobacterialproteomes. A total of 45 mycobacterialreference proteomes were present in UniProtKB in October, 2017. If a chokepoint protein showed ≥50% identity over 80% of sequence length in a minimum of 10 mycobacterialproteomes, it was considered as a part of the core proteome. (b) A chokepoint protein was considered as essential protein if it had an alignment identity ≥50% with a protein contained in Database of Essential Genes (DEG), over 80% of sequence length. Proteins which could not qualify either criterion were removed from further analysis.

The potential drugs which can block the IPPP were searched using DrugBank, the most widely used database of drug molecules [(Wishart et al. 2008)](https://paperpile.com/c/a3wH5Q/IRZIf). Currently DrugBank contains approximately 8200 different categories of drugs namely, US Food and Drug Administration (FDA)-approved small-molecules drugs, FDA-approved biotech drugs, nutraceuticals, and experimental drugs. To find the appropriate drug candidate, we downloaded sequences of all four types of targets namely, drug targets, drug enzymes, drug carriers and drug transporters from DrugBank database. Using BLAST we searched for homologs of chokepoint proteins among DrugBank target proteins. The drug molecule associated with the best hit of DrugBank target protein was considered as a potential binder of homologous chokepoint proteins. Here too, a hit was considered as a homologous protein if it showed ≥50% identity over 80% of the sequence length. In the next stage, these probable drugs were further optimized according to the Lipinski’s rule of five scales that includes molecular weight ≤500, number of rotatable bonds ≤10, hydrogen bond donors ≤5, hydrogen bond acceptors ≤10 and logP ≤5. Additionally, half-life ≥60 minutes and toxicity information were also considered while evaluating a drug molecule. Those drug molecules which possessed a minimum 5 of the 7 parameters were considered as probable drugs. Drug-like compounds categorized by the DrugBank as dietary supplements, micronutrients, or vitamins were excluded from the study.

## 5.2.5. Drug-target interaction

Molecular docking is a useful tool for modeling the interaction between two biomolecules or a small molecule (might be a drug) and a biomolecule at the atomic level. It allows us to model the behavior of binding partners in terms of binding affinity or interaction. In the present work we used PatchDock server to dock drug molecules [(Schneidman-Duhovny et al. 2005)](https://paperpile.com/c/a3wH5Q/1KSBE),which qualified the filtration criteria with their potential targets. The protein-ligand complex with the highest docked score was selected for further analysis. The structures of potential drugs were downloaded from DrugBank.

## 5.2.6. Drug repurposing for *M. tuberculosis*

Molecular mimicry plays an important role in primary establishment of *M. tuberculosis* inside the host. Hence, if *M. tuberculosis* mimicry inducing proteins can be blocked, the pathogen can be eliminated, way-before it establishes itself inside the host. The steps described above, can be also used to propose novel drugs against *M. tuberculosis*.

**5.3. Result and discussion**

## 5.3.1. Identification of interacting protein of mimicry proteins

Keyword search in miPepBase using “mycobacterium” displayed 25 entries/events related to mimicry (Table 5.1). In the 25 events, 20 distinct proteins of *Mycobacterium* involved in molecular mimicry were identified. These proteins were responsible for 7 different types of autoimmune diseases caused due to cross-reactivity with 12 different types of host proteins. We observed that one protein of the pathogen (A0A040DMG3) was removed by UniProt; hence it was excluded from our further studies. The 7 different types of autoimmune diseases caused by the remaining 24 molecular mimicry events were encephalomyelitis, leprosy, multiple sclerosis, primary biliary cirrhosis, rheumatoid arthritis, skin disease and type 1 diabetes (Table 5.1). Also, not all of the 24 molecular mimicry events were caused due to the proteins of *M. tuberculosis*, 1 event was due to proteins of *M. avium,* 6 due to proteins of *M. avium* subsp. *paratuberculosis,* 4 due to proteins of *M. leprae,* 1 due to proteins of *M. gordonae,* 11 due to proteins of *M. tuberculosis* and 1 due to proteins of *M. bovis.* Using STRING, of the 19 path-proteins, we were able to find interacting partners for 16 proteins. Among the 16 path-proteins, one protein (P9WQ90) was a homo-dimer while two proteins (P0A521 and Q49375) were oligomers. For those path-proteins (A0A045I964, A0A0E2WUC4 and Q53467), about which protein-interaction information could not be retrieved using STRING [11], BLAST search against the UniProtKB database was used to find homologous proteins. The first hits retrieved after BLAST search of A0A045I964 and A0A0E2WUC4 were I6XH73 and F5Z390, respectively. However, for path-protein Q53467 we did not find any hit with high sequence homology (Table E5.2). Hence, it was removed from further analysis. Both I6XH73 and F5Z390 were also mycobacterial proteins. STRING search revealed that for the 15 path-proteins, there were 148 interacting protein partners. In the present work, if IPPP had alignment identity less than 50% with alignment coverage less than 80% with a human protein, it was considered as non-homologous IPPP (nHIPPP). As per this guideline among 148 IPPP, 5 proteins were homologous IPPP. Hence, these were also excluded from further analysis (Table 5.2).

**5.3.2. Identification and validation of the metabolic chokepoints**

The number of pathways to which these proteins mapped was ̶ 12 for *M. leprae,* 14 for *M. avium* subsp. *paratuberculosis* and 18 for *M. tuberculosis*. The pathways were analyzed manually to find possible chokepoint reaction(s). Our analysis revealed that these 143 proteins were a part of 53 chokepoints. We found that 47 of the 53 choke-points proteins were part of core proteins (Table E5.3), and 31 of the 53 chokepoint proteins shared a close homolog with DEG proteins (Table 5.3).

**Table 5.1.** List of *Mycobacterium. spp*. protein that are involved in molecular mimicry (source: Garg, A., et al. (2017) miPepBase: A Database of Experimentally Verified Peptides Involved in Molecular Mimicry. *Front Microbiol*. 8, 2053).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Pathogen protein entry (UniProt AC)** | **Mimicry peptide** | **Pathogen protein name** | **Pathogen name** | **Host Name** | **Host Protein Entry (UniProt AC)** | **Host Protein Name** | **Host mimicry peptide** | **Autoimmune disease** |
|  | A0A040DMG3\* | ACFTRPARWTL | Transmembrane protein | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | A0A045I964 | QRCRVHFMRNLYTAV | Transposase | *M. tuberculosis* | Human | P02686 | MBP | ENPVVHFFKNIVTPR | Multiple sclerosis |
|  | A0A0E2WUC4 | QRCRVHFLRNVLAQV | Transposase | *M. avium* | Human | P02686 | Myelin basic protein (MBP) | ENPVVHFFKNIVTPR | Multiple sclerosis |
|  | A5U2C2 | AAQHRQIVADF | UvrABC system protein C | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | A5U956 | AAQARPVKTVI | MYCTX Transferase | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | O32984 | VSPWGKPEGRTRKPNKSSNK | 50S ribosomal L2 | *M. leprae* | Mice | P02687 | Myelin basic protein | VVHFFKNIVTPRTPPPSQGK | Leprosy |
|  | O32984 | EQANINWGKAGRMRWKGKRP | 50S ribosomal L2 | *M. leprae* | Mice | P02687 | Myelin basic protein | GAPKRGSGKDGHHAARTTHY | Leprosy |
|  | P09239 | NA | 65-kDa heat shock protein | *M. leprae* | Human | P13645 | Cytokeratin-10 of keratin | NA | Leprosy |
|  | P0A521 | AGKPLLIIAEDVEGE | HSP65 | *M. bovis* | Human | P10809 | HSP60 | HRKPLVIIAEDVDGE | Rheumatoid arthritis |
|  | P46861 | NTLSAPTFVKDFPVETTPLT | Lysyl- tRNA synthetase | *M. leprae* | Mice | P02687 | Myelin basic protein | VVHFFKNIVTPRTPPPSQGK | Leprosy |
|  | P9WG07 | AYYGALPLIV | ABC transport | *M. tuberculosis* | Rabbit | P25274 | Mid-region encephalitogen from myelin basic protein | TTHYGSLPQK | Multiple Sclerosis |
|  | P9WM57 | ATQYRPDQLAK | Uncharacterized protein R | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | P9WN15 | ASMNRPNLVAL | Uncharacterized glycosyl hydrolase | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | P9WPE7 | STVKDLLPLL | 65-kDa heat shock protein | *M. tuberculosis* | Rat | P02788 | Human lactoferrin | SGQKDLLFKD | Rheumatoid arthritis |
|  | P9WPE7 | STVKDLLPLL | 65-kDa heat shock protein | *M. tuberculosis* | Rat | P02787 | Human transferrin | PHGKDLLFKD | Rheumatoid arthritis |
|  | P9WPE7 | VPGGGDMGG | 65-kDa heat shock protein | *M. tuberculosis* | Human | P12035 | Human keratin | GGYGGGMGG | Skin diseases |
|  | P9WPE7 | VPGGGDMGG | 65-kDa heat shock protein | *M. tuberculosis* | Human | P10809 | Human hsp65 | GGMGGGMGG | Skin diseases |
|  | P9WQ90 | ASHQRQRAFAQ | Probable aspartate aminotransferase | *M. tuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | Q49375 | GDL(IL)AE | 65-kDa heat shock protein | *M. gordonae* | Human | P10515 | Pyruvate dehydrogenase complex-E2 | GDLIAE | Primary Biliary Cirrhosis |
|  | Q53467 | SHQIRPVCGQR | Putative transport protein | *M. avium subsp. paratuberculosis* | Mouse | F6RT34 | Myelin basic protein MBPAc | ASQKRPSQRSK | Encephalomyelitis |
|  | Q73T54 | MIAVALAGL | Uncharacterized protein | *M. avium subsp. paratuberculosis* | Human | Q8IWU4 | Beta-cell protein zinc transporter 8 (ZnT8) | MIIVSSCAV | Type 1 diabetes |
|  | Q73T54 | LAANFVVAL | Uncharacterized protein | *M. avium subsp. paratuberculosis* | Human | Q8IWU4 | Beta-cell protein zinc transporter 8 (ZnT8) | VAANIVLTV | Type 1 diabetes |
|  | Q73WP1 | WYIPPLSPVV | MAP\_2619 | *M. avium subsp. paratuberculosis* | Human | Q16653 | Human myelin oligodendrocyte glycoprotein | MEVGWYRPPFSRVVHLYRNGK | Multiple Sclerosis |
|  | Q741P6 | LKYGSLPLSF | SecD | *M. avium subsp. paratuberculosis* | Rabbit | P25274 | Mid-region encephalitogen from myelin basic protein | TTHYGSLPQK | Multiple Sclerosis |
|  | Q745A5 | PGRRPFTRKELQ | Uncharacterized protein | *M. avium subsp. paratuberculosis* | Human | P02686 | Myelin basic protein | ENPVVNFFKNIVTP | Multiple sclerosis |

\* Shows obsolete UniProtKB entry

**Table 5.2. List of pathogen mimicry protein, its interaction partners (IPPP), name of human homolog (if present), KEGG pathway ID to which IPPP belongs and chokepoint proteins.** The table shows information related to pathogen protein involved in molecular mimicry (Column 2), IPPP collected from STRING database at default parameters (Column 3), HIPPP among the IPPP (column 4), IPPP which couldn’t be mapped on KEGG (column 5), KEGG pathway ID in which IPPP mapped (Column 6) and chokepoints proteins found after manual survey of KEGG pathways ID listed in column 6 (Column 7).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.No.** | **Path- proteins** | **Interacting proteins of pathogen’s proteins (IPPP)** | **Human Homolog of IPPP** | **Proteins couldn’t be mapped on KEGG** | **KEGG Pathway ID** | **Chokepoint proteins** |
| 1 | A5U2C2 | MRA\_1028, MRA\_1424, MRA\_1430, MRA\_1431, MRA\_1432, MRA\_1648, MtubH3\_010100010416, mfd, uvrA, uvrB, uvrC | NA | MRA\_1028, MRA\_1424, MRA\_1430, MRA\_1431, MRA\_1432, MRA\_1648, MtubH3\_010100010416 | mtu03420 | Mfd, uvrA, uvrB, uvrC |
| 2 | A5U956 | MRA\_0226, MRA\_0381, MRA\_1886, MRA\_3014, MRA\_3766, MRA\_3767, MRA\_3895, ethR | NA | NA | NA | NA |
| 3 | F5Z390 | JDM601\_3772, JDM601\_3773, JDM601\_3774 | NA | NA | NA | NA |
| 4 | I6XH73 | Rv3431c, gadB, nnr, Rv3434c, Rv3435c | NA | Rv3431c, nnr, Rv3434c, Rv3435c | mtu00410, mtu01100, mtu00650, mtu01120, mtu00250, mtu00430, mtu02024, mtu01110 | mtu02024: gadB |
| 5 | O32984 | rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS | NA | NA | mle03010 | rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS |
| 7 | P09239 | clpB, dnaJ1, dnaJ2, dnaK, groL2, groS, grpE, hrcA, htpG, mdh, pheT | dnaK,  mdh | clpB, dnaJ1, dnaJ2  , groS, grpE, hrcA, htpG | mle03018, mle05152, mle00970 | mle03018:groEL; mle05152:groEL; mle00970:pheT |
| 6 | P0A521 | NA | NA | NA | NA | NA |
| 8 | P46861 | argS, gltX, guaA, ileS, leuS, lysS, lysX (ML1393 ), metG, panC, pheT, proS | NA | NA | mle00970, mle01100, mle01110, mle00770, mle00230, mle00860, mle00450, mle00410, mle01120 | mle00970:gltX,metG,leuS,ileS,lysS,argS,proS,pheT; mle00860:gltX; mle00410:panC |
| 9 | P9WG07 | phoU1 (Rv3301c ), phoU2 (Rv0821c), pstA1, pstA2, pstB1 or phoT (Rv0820 ), pstB2 or pstB (Rv0933), pstC1, pstS1, pstS2, pstS3, tcrX (Rv3765c ) | NA | phoU1 (Rv3301c ), phoU2 (Rv0821c), tcrX (Rv3765c ) | mtu02010, mtu02020, mtu05152 | mtu02010:pstB1,pstB2,pstA1,pstA2,pstC1,pstS1,pstS2,pstS3; mtu02020:pstS1,pstS2,pstS3; mtu05152:pstS1,pstS2,pstS3 |
| 10 | P9WM57 | Rv0184, Rv0336, Rv0515, Rv1128c, Rv1278, Rv1378c, Rv1765c, Rv2015c, Rv2100, Rv3074, Rv3776 | NA | NA | NA | NA |
| 11 | P9WN15 | Rv2006, Rv3400, Rv3401, aglA, glgB, glgE, glgX (Rv1564c ), otsA, otsB (Rv3372), treS, treZ | NA | Rv3400, Rv3401 | mtu01100, mtu00500, mtu01110, mtu00052 | Mtu00500: glgB; mtu00052: aglA |
| 12 | P9WPE7 | Rv0312, Rv2264c, dnaJ1, dnaJ2, dnaK, groL2 (Rv0440), groS, hycE, metK, pheT, thrS | dnaK,  metK | Rv0312, Rv2264c, dnaJ1, dnaJ2, groS, hycE | mtu00970, mtu05152, mtu03018 | Mtu00970: thrS,pheT; mtu05152:groL2; mtu03018:groL2 |
| 13 | P9WQ90 | NA | NA | NA | NA | NA |
| 14 | Q49375 | NA | NA | NA | NA | NA |
| 15 | Q53467 | NA | NA | NA | NA | NA |
| 16 | Q73T54 | MAP\_2073c, MAP\_2138, MAP\_2784, MAP\_2925, MAP\_3865c, MAP\_3866c, MAP\_3867c, atpA, ctpA, ctpC, nrdE | atpA | MAP\_2073MAP\_2138, MAP\_2784, MAP\_2925, MAP\_3865c, MAP\_3866c, MAP\_3867c, ctpA, ctpC | mpa00230, mpa00240, mpa00190 | NA |
| 17 | Q73WP1 | MAP\_0368, MAP\_2102c (narK3\_1), MAP\_3636, MAP\_3707c (narK3\_2), MAP\_4101c, fdhF, narG, narH, narI, narJ, narU | NA | MAP\_3636, MAP\_4101c, fdhF | mpa00910, mpa01120, mpa02020, mpa01100, mpa00630, mpa00680, mpa01200 | mpa00910: narK3\_1, narK3\_2, narU, narH, narG |
| 18 | Q741P6 | MAP\_1042, MAP\_1045, apt, dnaG, relA, secD, secE, secF, secG, secY, tatC | NA | NA | mpa03070, mpa03060, mpa02024, mpa00230, mpa01100, mpa03030 | Mpa03070: MAP\_1042, secD, secE, secF, secG, secY; mpa03060:MAP\_1042, secD, secE, secF, secG, secY; mpa02024:MAP\_1045,MAP\_1042,secE,secG,secY; mpa03030:dnaG |
| 19 | Q745A5 | MAP\_0105c, MAP\_0106c, MAP\_1410, MAP\_2148, MAP\_2752, MAP\_2963c, MAP\_3314c, ftsK, ogt, parB, topA | NA | NA | NA | NA |

## 

## 5.3.3. Drug molecules for metabolic chokepoints

Using all DrugBank target and chokepoint protein pairs, for the 5 potential chokepoint proteins, we were able to identify 11 drug candidates (Table E5.4). Proteins against which we could find drugs were mostly interaction partners of mimicry proteins responsible for multiple sclerosis. After benchmarking on the basis of Lipinski’s rule along with toxicity and half-life of drug molecules, we were finally left with 4 probable drug candidates. Of these 4 probable drug candidates, we noted that 3 were experimental approved drugs *viz.* DB08185, DB00759 and DB01930 against three chokepoints proteins of *M. leprae* rpsS, rpsC and panC, respectively (Table 3). The fourth drug candidate was an experimentally verified drug, DB07349 against narH of *M. avium* subsp. *paratuberculosis*.

Of the four drugs, DB01930 is known to target the enzyme pantothenate synthetase of *M. tuberculosis* (<https://www.drugbank.ca/drugs/DB01930>). Pantothenate synthetase catalyzes the ATP-dependent condensation of pantoate and β-alanine to form pantothenate (vitamin B5) [(Jonczyk et al. 2008)](https://paperpile.com/c/a3wH5Q/jDzc6). It is a known fact that pantothenate biosynthesis is essential for virulence of *M. tuberculosis* [(Hung et al. 2016)](https://paperpile.com/c/a3wH5Q/uqirb). DB07349 is an experimental drug, which targets narH and L. Several *in vivo* studies indicate that human lung granuloma where *M. tuberculosis* resides during latency, are hypoxic and narH and L plays an important role in bacterial survival in hypoxic environment. This suggests that DB07349 can be an ideal drug candidate because it can kill *M. tuberculosis* residing in granuloma. Also, DB07349 can help in complete clearance of *Mtb* from the host, because long-term persistence of *M. tuberculosis* in the latent stage not only helps it in remaining unaffected during the anti-tubercular treatment but also help the pathogen to develop resistance against currently used drugs [(Gomez & McKinney 2004)](https://paperpile.com/c/a3wH5Q/mHyPl). The targets of other two drugs DB00759 & DB08185 are parts of ribosomal protein complex. DB00759 (commonly known as Tetracycline) is already an approved drug. It is being given to patients through oral route as well as an ophthalmic ointment. These are also reported to inhibit the *Mtb* pathogen growth by binding to the 30S ribosomal sub-unit and blocking translation [(Hentschel et al. 2017)](https://paperpile.com/c/a3wH5Q/PcML0). We observed that among all chokepoint proteins 3D structure of only panC was available in PDB. Hence the 3D structures of remaining proteins were obtained from Swiss-model (rpsS and rpsC) and modBase (narH). The intermolecular interactions and strengths, hydrogen bonding, hydrophobic interactions and atom accessibilities are shown in Table E5.5.

## 5.3.4. Drug repurposing for *M. tuberculosis*

The fifty-three chokepoints proteins identified, belonged to three different species of mycobacteria. Hence,their homologs were searched in the proteomeof *M. tuberculosis*. We observed that, of the fifty-three chokepoints (14 of *M. tuberculosis*, 20 of *M. leprae* and 13 of *M. avium* subsp. *paratuberculosis*) homologous proteins for forty-seven chokepoint proteins were present in the proteome of *M. tuberculosis* (Table 5.3). Hence, we anticipate that these four drugs *viz.* DB08185, DB00759, DB01930 and DB07349 might be useful in treatment of *M. tuberculosis*.

**Table 5.3**. **Drug target validation.** The table shows information of path-proteins (Column 1), potential chokepoint found in KEGG metabolic network (Column 2), chokepoint proteins which were part of essential genes (Column 3) and core proteins (Column 4), homologous of chokepoint proteins in Mtb proteome (Column 5) and chokepoint protein listed as drug target in DrugBank database (Column 6). Column 7 has potential drug molecule as per DrugBank target protein and Column 8 contained the drugs, which qualified the filter of drug candidate filter.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Path-protein** | **Chokepoint proteins** | **Chokepoint proteins which were part of Essential gene database** | **Chokepoint proteins which were part of Core proteome** | **Homologue of chokepoint proteins in Mtb proteome** | **Chokepoint proteins included as drug-target in DrugBank** | **Potential Drug molecule as per DrugBank against target proteins** | **Drugs follow at least 5 of 7 drug like properties** |
| A5U2C2 | mfd, uvrA, uvrB, uvrC | uvrC | mfd, uvrA, uvrB, uvrC | mfd, uvrA, uvrB, uvrC | NA | NA | -- |
| I6XH73 | gadB | NA | gadB | gadB | gadB | gadB: DB03553 | -- |
| O32984 | rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS | rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS | rplB, rplC, rplD, rplF, rplN, rplP, rplW, rplV rpmC, rpsC, rpsS | rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS | rpsC, rpsS | rpsC: DB00759; DB09093  rpsS: DB08185; DB00560; DB00759; DB09093 | DB08185 (2-Methylthio-N6-Isopentenyl-Adenosine-5'-Monophosphate), DB00759 (Tetracycline) |
| P09239 | mle03018: groEL; mle05152:groEL; mle00970:pheT | groL2, pheT | groL2, pheT | groL2, pheT | NA | NA | -- |
| P46861 | mle00970: gltX,metG,leuS,ileS,lysS,argS,proS,pheT; mle00860:gltX; mle00410:panC | gltX, metG, leuS, ileS, lysS, argS, pheT, panC | gltX, metG, leuS, ileS, lysS, argS, pheT, panC | gltX, metG, leuS, ileS, lysS, argS, pheT, panC | panC | panC: DB01930; DB02596; DB02694; DB03107 | DB01930 ((1S)-2-{[{[(2S)-2,3-Dihydroxypropyl]Oxy}(Hydroxy)Phosphoryl]Oxy}-1-[(Pentanoyloxy)Methyl]Ethyl Octanoate) |
| P9WG07 | mtu02010: pstA1,pstA2,pstB1,pstB2, pstC1,pstS1,pstS2,pstS3; mtu02020:pstS1,pstS2,pstS3; mtu05152:pstS1,pstS2,pstS3 | NA | pstA1, pstB1,pstS2, pstS3 | pstA1, pstB1,pstS2, pstS3 | NA | NA | -- |
| P9WN15 | Mtu00500: glgB; mtu00052: aglA | glgB | glgB, aglA | glgB, aglA | NA | NA | -- |
| P9WPE7 | Mtu00970: thrS,pheT; mtu05152:groL2; mtu03018:groL2 | thrS, pheT, groL2 | thrS, pheT, groL2 | thrS, pheT, groL2 | NA | NA | -- |
| Q73WP1 | mpa00910: narK3\_1, narK3\_2, narU, narH, narG | NA | narK3\_1, narK3\_2, narH, narG | narK3\_1, narK3\_2, narU, narH, narG | narH | narH: DB04464; DB07349 | DB07349 (2,4-Dihydroxy-3,3-Dimethyl-Butyrate) |
| Q741P6 | Mpa03070:MAP\_1042, secD, secE, secF, secG, secY; mpa03060:MAP\_1042, secD, secE, secF, secG, secY; mpa02024:MAP\_1045,MAP\_1042,secE,secG,secY; mpa03030:dnaG | secD, secE, secF, secG, secY, dnaG | MAP\_1042, MAP\_1045, dnaG, secD, secE, secF, secG, secY | MAP\_1042, MAP\_1045, dnaG, secD, secE, secF, secG, secY | NA | NA | -- |

**5.4. Prospects of the current approach**

Much research has been done to discover novel drug targets and potent drugs against TB [(Anishetty et al. 2005, Duncan 2004, Ioerger et al. 2013, Lamichhane 2011, Mdluli & Spigelman 2006, Raman et al. 2008, Singh & Mizrahi 2017)](https://paperpile.com/c/a3wH5Q/lMYtJ+oeQAL+mYrAI+OfhCO+vfXtg+jufHw+lNL7O). The current approach is different from earlier approaches, because here our target is not an active physiological process/protein(s) which helps in establishing TB bacteria inside the host. Our target is protein(s) (and their interacting partners) which are responsible for eliciting auto-immunity inside the host. In the current study, the authors propose to target/disrupt proteins of *M. tuberculosis* which evoke autoimmune diseases (using drugs/chemical compounds) as a prophylactic measure, before the onset of active TB infection. It would be pertinent to mention here that a recent research proposed that mycobacterial infections might have driven autoimmunity as an evolutionary strategy and proteins involved in molecular mimicry are produced in the host long before the appearance of symptoms of TB [(Elkington et al. 2016)](https://paperpile.com/c/a3wH5Q/HZU7I). Thus, our approach might be useful in devising novel prophylactic/vaccination measures against TB.

Another prospective use of our approach is that it can be used as a follow up remedy after a patient is cured from TB. The drug molecules identified in our current study would disrupt the growth of latent bacteria residing inside the host, which will ultimately lead to clearance of TB bacilli from host. The other advantage of our approach is that it is in-line with the therapy used for treatment of autoimmune diseases. Tumor necrosis factor (TNF) blocker therapy is an effective treatment for many autoimmune diseases but it also significantly increases the risk of progression of latent TB to active TB. Thus, before commencing the TNF-blocker therapy for curing autoimmune diseases, patients are first tested for TB infection. Hence use of a drug, which does not involve the use of a TNF-blocker, can lead to significant improvement in treatment of pathogen induced auto-immunity.

**5.5. Concluding remarks**

Computational methods and integrated ‘omics’ approach, encompassing genomics, proteomics, and metabolomics have proved as a valuable tool in drug-discovery. Comparative and subtractive genomics proved helpful for prediction and identification of potential therapeutic targets and vaccine candidate proteins in numerous pathogenic bacteria and fungi [(Amineni et al. 2010, Chong et al. 2006, Doyle et al. 2010, Johri et al. 2006)](https://paperpile.com/c/a3wH5Q/mlkaN+Aka8R+UaM2C+1nKXz) In the current review, we have proposed novel drug targets and drug molecules using pathogen’s molecular mimicry inducing proteins. The identification has been done by employing a rigorous systems biology approach. The process and the workflow for identification of drug targets have been explained in detail using *M. tuberculosis,* as the model organism. Our systematic analysis revealed that interacting proteins of mimicry inducing proteins of mycobacteria contain several chokepoints proteins, which can serve as potential drug targets. Inhibitors of the choke point proteins were searched from the DrugBank employing several stringent filters. The DrugBank search revealed three drug compounds enlisted in the experimental group and, one in approved group, which might be effective against *M. tuberculosis*. Interaction between target(s) and their cognate drug molecule(s) was further confirmed by molecular docking. The drug candidates identified during the course of this study are FDA approved drug-molecules, with proven efficacy against many microbial pathogens. These drug candidates are propose to be tested in laboratory *in vitro* conditions for assessing their efficacy against *M. tuberculosis* clinical isolates. Thus, instead of developing new chemotherapeutics our approach helps in repurposing the known drugs against TB.

Using the interaction partners of mimicry proteins, the authors were able to discover four drug-candidates against TB. The trivial number of drug might be due to the fact that only one database was used to search drug molecules, the DrugBank. DrugBank was preferred over other databases because it provides detailed information about the properties and mechanism of action of approximately 12000 marketed or experimental drugs. However, the number of probable drug candidates would have increased if data from other relevant databases were also included in the study. For example databases, such as ChEMBL [(Gaulton et al. 2012)](https://paperpile.com/c/a3wH5Q/eenA1), PubChem [(Kim et al. 2016)](https://paperpile.com/c/a3wH5Q/XVirO), and ChemBank [(Seiler et al. 2008)](https://paperpile.com/c/a3wH5Q/188CV) could be used to provide comprehensive collection of biological activity where as ZINC database [(Irwin & Shoichet 2005)](https://paperpile.com/c/a3wH5Q/gqQ8m) could be used for virtual screening. Similarly incorporation of additional data of protein–chemical interactions such as Therapeutic Target Database (TTD) [(Li et al. 2018)](https://paperpile.com/c/a3wH5Q/PNEg3) and STITCH [(Szklarczyk et al. 2016)](https://paperpile.com/c/a3wH5Q/yq8X7)) can also increase number of drug targets and candidates.

Nevertheless, authors believe that the scheme described in the current study can be applied for repurposing the known drugs and discovery of novel therapeutics against other pathogenic bacteria, which exhibit molecular mimicry with the host’s proteins.

The results described in this chapter have been published as - Garg *et al.* Using molecular-mimicry-inducing pathways of pathogens as novel drug targets. ***Drug discovery today***, 24, 9, 1943-1952, 2019

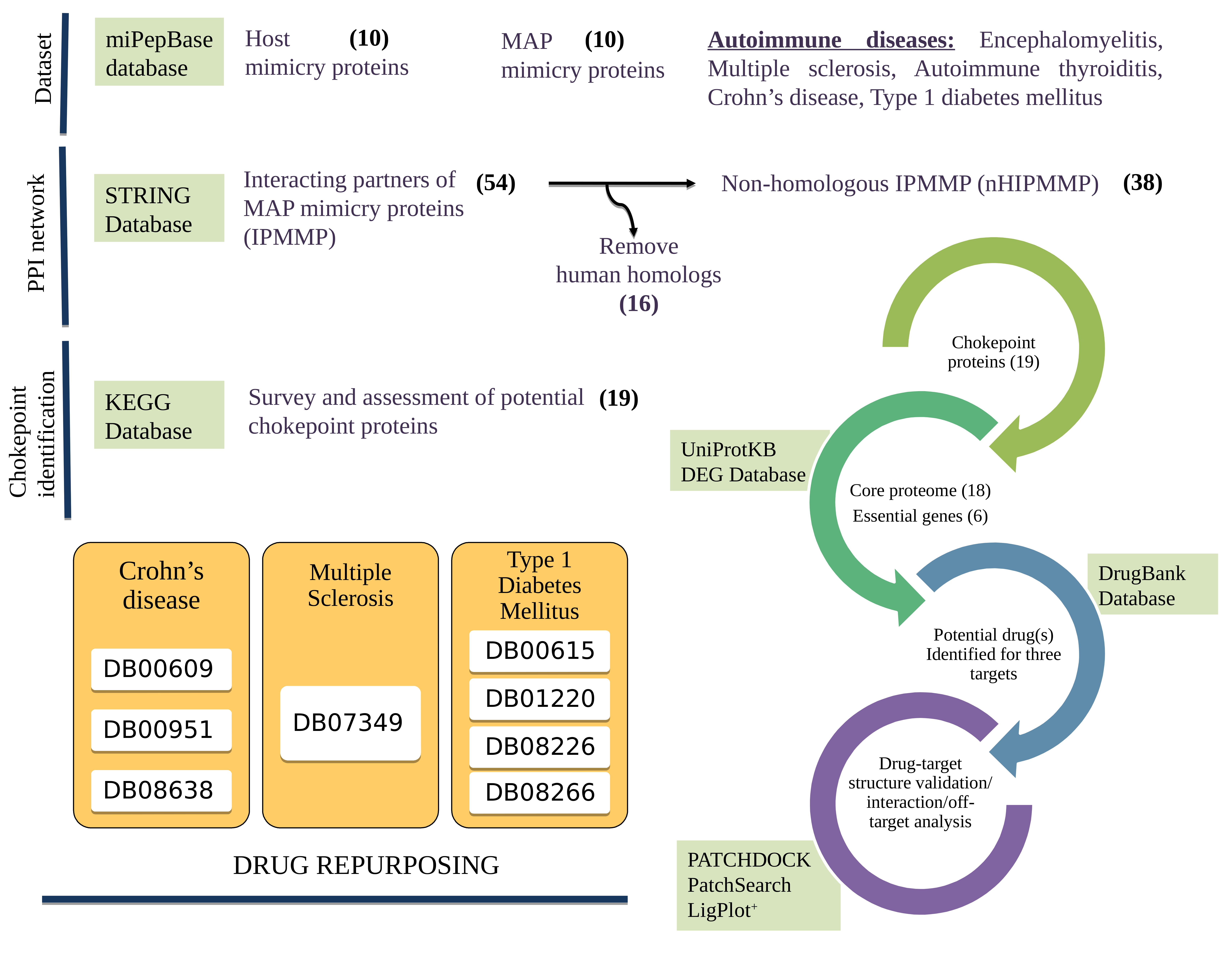
**Chapter 6**

# 6.1. Introduction

(MAP) belongs to the *Mycobacterium avium* complex which contains mycobacterial species like *M. avium* and *M. intracellulare* [(Shin et al. 2010)](https://paperpile.com/c/a3wH5Q/gePeU). It is an obligate intracellular bacterial pathogen that can cause chronic granulomatous gastroenteritis in ruminants, called Johne's disease [(Dow & Ellingson 2010, Sechi & Dow 2015)](https://paperpile.com/c/a3wH5Q/xtUd5+3hyI5). Humans might become infected with MAP through fecal-oral, waterborne-, foodborne- and zoonotic-routes [(Robertson et al. 2017, Whiley et al. 2012, Wynne et al. 2011)](https://paperpile.com/c/a3wH5Q/6ZKR5+UePoy+rbgga). MAP exhibits ‘epitope mimicry’ with the host peptide(s)/protein(s), which elicits host autoreactive T- or B-cells leading to tissue and/or organ damage and, ultimately autoimmune diseases [(Bitti et al. 2012)](https://paperpile.com/c/a3wH5Q/YdXqh). It is widely believed that autoimmune diseases develop in genetically predisposed individuals due to some environmental triggers [(Dow 2012, Miller 2011)](https://paperpile.com/c/a3wH5Q/br8Ex+dfBKf). Host genetic factors like polymorphisms in the CARD15 (NOD2) [(Dow & Ellingson 2010, Sechi & Dow 2015, Sechi et al. 2005a)](https://paperpile.com/c/a3wH5Q/xtUd5+3hyI5+iSKfT), SLC11a1 (NRAMP1) [(Dow 2012, Sechi et al. 2006, Wang et al. 2018)](https://paperpile.com/c/a3wH5Q/JkYvs+dfBKf+R70go), LRRK2 [(Dow 2014, Härtlova et al. 2018)](https://paperpile.com/c/a3wH5Q/gFV1F+f87Pq), PTPN2/22 [(Sharp et al. 2018)](https://paperpile.com/c/a3wH5Q/GG2GH), and VDR [(Dow 2006)](https://paperpile.com/c/a3wH5Q/wHNaH) genes have been associated with MAP and various diseases like Crohn’s disease (CD) [(Sechi & Dow 2015, Sechi et al. 2006)](https://paperpile.com/c/a3wH5Q/R70go+xtUd5), Blau syndrome [(Sechi & Dow 2015)](https://paperpile.com/c/a3wH5Q/xtUd5), multiple sclerosis (MS) [(Cossu et al. 2013, Sechi & Dow 2015)](https://paperpile.com/c/a3wH5Q/xtUd5+OodZQ), Hashimoto’s thyroiditis [(D’Amore et al. 2010, Gong et al. 2018, Sisto et al. 2010)](https://paperpile.com/c/a3wH5Q/VFr3x+4kIa2+ToAtc), Parkinson’s disease [(Arru et al. 2016, Dow 2014)](https://paperpile.com/c/a3wH5Q/rwmbu+gFV1F), rheumatoid arthritis [(Bo et al. 2018, Sharp et al. 2018, Wang et al. 2018)](https://paperpile.com/c/a3wH5Q/GG2GH+JkYvs+5z9aL) and type 1 diabetes mellitus (T1DM) [(Dow 2006, Paccagnini et al. 2009)](https://paperpile.com/c/a3wH5Q/EgzaA+wHNaH). Several studies have suggested MAP as an environmental trigger for multiple human autoimmune diseases like MS, T1DM, Hashimoto's thyroiditis, sarcoidosis, CD, and Blau syndrome [(D’Amore et al. 2010, Mameli et al. 2016, Naser et al. 2004, Quayle et al. 1992, Robertson et al. 2017, Sechi et al. 2005b)](https://paperpile.com/c/a3wH5Q/VFr3x+rbgga+umP2Y+tiJ66+XAkIe+Vi7Qz). MAP was reportedly also detected from the body samples of patients suffering from autoimmune T1DM and MS patients [(Chandrashekara 2012, Dow 2012, Mameli et al. 2014, Miller 2011)](https://paperpile.com/c/a3wH5Q/br8Ex+dfBKf+gJrjO+uAquW)

The conventional therapy for autoimmune diseases has been the usage of immune-suppressants or immune-modulators which treat symptoms rather than the etiology and/or the causative mechanism(s). Though, 60–70% of the patients initially respond to immunosuppression, in many cases the patients show subsequent clinical remission or relapse of the autoimmune disease [(Chandrashekara 2012, van der Kooij et al. 2007)](https://paperpile.com/c/a3wH5Q/37qW9+gJrjO). Studies have indicated that immune-suppressants conventionally used for treating MAP-associated autoimmune diseases actually inhibited the growth of MAP *in vitro* [(Greenstein et al. 2008, 2010)](https://paperpile.com/c/a3wH5Q/KmcP7+j1CSy). Interestingly, the drug regimen consisting of clarithromycin, rifabutin, and clofazimine (anti-mycobacterial drugs) was reportedly effective in the treatment of MAP-associated autoimmune diseases like CD [(Chamberlin et al. 2011)](https://paperpile.com/c/a3wH5Q/DkDab) and MS (https://clinicaltrials.gov/ct2/show/NCT01717664). This suggests that eliminating the etiopathological agent itself might be a better strategy for treating MAP-associated autoimmune diseases.

In the present study, we have performed a systematic *in silico* analysis of metabolic chokepoints of MAP mimicry proteins and their interacting protein partners to identify suitable drug targets which can eliminate the pathogen itself and, thereby treat various MAP-associated autoimmune diseases. Metabolic pathway/metabolic network analysis and identification of metabolic chokepoints are widely used *in silico* methods to identify drug targets in pathogen genomes. By definition, a chokepoint enzyme either consumes a unique substrate or produces a unique product in the pathogen metabolic network [(Yeh et al. 2004)](https://paperpile.com/c/a3wH5Q/hXqnX). Inhibition of chokepoint enzymes may disrupt crucial metabolic processes in the pathogen, so chokepoints that are essential to the pathogen represent good potential drug targets [(Chung et al. 2013, Duffield et al. 2010, Garg et al. 2019, Sharma & Pan 2012, Siwo et al. 2015)](https://paperpile.com/c/a3wH5Q/KuHvh+sjg4H+aHc8P+A9jFn+CJqgW). In our study, information regarding the interacting proteins of MAP involved in molecular mimicry with the host proteins was retrieved from the STRING database. All the interacting partners of MAP mimicry proteins which showed homology with the human proteins were removed, and only non-homologous MAP proteins were included in further analysis. Then we determined the chokepoint(s) of the metabolic pathways, followed by further confirmation of the essentiality of the identified chokepoints in the survival of MAP. Finally, the potential drugs which can block the chokepoint(s) were identified using the DrugBank database and confirmed by molecular docking (Figure 6.1). Additionally, the “off-target” binding potential of each proposed drug was also assessed.



**Figure 6.1.** The workflow adopted for the identification of novel drug targets for treating *M. avium* subsp*. paratuberculosis*-associated autoimmune disorders.

# 6.2. Material and Methods

## 6.2.1. Retrieval of experimentally validated MAP mimicry proteins involved in autoimmune diseases

The information about MAP proteins involved in molecular mimicry and autoimmune diseases was retrieved from the database - miPepBase [(Garg et al. 2017)](https://paperpile.com/c/a3wH5Q/ElOF5). The miPepBase is a database of experimentally validated mimicry proteins which was earlier developed in our laboratory. It contains extensive manually curated information about all the mimicry proteins/peptides, and the autoimmune diseases reported, till date. Here, we used "*Mycobacterium avium* subsp. *paratuberculosis*” as the keyword to access the data.

## 6.2.2. Protein-protein interaction studies

The STRING database was used to retrieve the information about the interacting partners of MAP mimicry proteins (IPMMP). STRING contains information about protein-protein interactions (PPI) which were established by experimental studies and, also by different methods of genome analysis like domain fusion, phylogenetic profiling, and gene neighborhood [(Mering & v. Mering 2003)](https://paperpile.com/c/a3wH5Q/Z65oj). Each PPI of STRING is assigned with a confidence score (0.15 = low confidence, ≥0.4 = medium confidence and default threshold, ≥0.7 = high confidence, and ≥0.9 = highest confidence). The STRING confidence score is a combined score of eight methods namely neighborhood on the chromosome, gene fusion, phylogenetic co-occurrence, homology, coexpression, experimentally determined expression, database annotated, and automated text mining. If the combined STRING score is >0.5 then the chances of false-positive interactions in the second shell are quite less [(Szklarczyk et al. 2017)](https://paperpile.com/c/a3wH5Q/5Kkbq). Hence, in the present work, we have used a high confidence STRING combined score threshold of ≥0.7 to reduce the false-positive and negative PPIs.

## 6.2.3. Removal of human homologs of MAP proteins

To avoid non-specific binding of a ligand to host proteins, MAP proteins which were homologous to human proteins were removed from the list of interacting proteins. In this study, if an IPMMP showed ≥30% identity and ≥80% coverage with a human protein, it was considered as homologous to the human protein and referred as homologous interacting partners of MAP mimicry proteins (HIPMMP), whereas remaining proteins were considered as non-homologous interacting partners of MAP mimicry proteins (nHIPMMP).

## 6.2.4. Determining the chokepoint(s) of the metabolic pathways of MAP

Each nHIPMMP was mapped in their corresponding metabolic networks in the KEGG database [(Kanehisa et al. 2017)](https://paperpile.com/c/a3wH5Q/jChxW). KEGG is a database resource that cross-integrates genomic, chemical, and systemic functional information of an organism. It is widely used as a reference knowledge base for the integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies. In the present work, each nHIPMMP was manually surveyed to assess their capability to be a potential chokepoint protein.

## 6.2.5. Validation of the essentiality of chokepoint proteins

The validation of the essentiality of chokepoint proteins in MAP metabolic pathways was performed using two parameters. First, we assessed the presence of homologs of the chokepoint proteins in all the mycobacterial reference proteomes present in UniProtKB (total 44 mycobacterial proteomes, as in December 2019) [(Apweiler et al. 2004)](https://paperpile.com/c/a3wH5Q/3tfSK). If a chokepoint protein exhibited ≥50% identity over 80% of the sequence length in at least 10 mycobacterial proteomes, it was considered as the part of the core proteome. Second, a chokepoint protein was considered as an essential protein if it showed similarity with other mycobacterial essential genes enlisted in the Database of Essential Genes (DEG) [(Zhang et al. 2004)](https://paperpile.com/c/a3wH5Q/x1jOF). The chokepoint protein(s) was considered an essential gene of MAP, if the alignment identity and sequence coverage with a protein in DEG was more than equal to 50% and 80%, respectively.

## 6.2.6. Identification and validation of drug molecules against the chokepoint proteins

The potential drugs which can block the chokepoint(s) were searched in the DrugBank, the most popular and well curated and regularly updated database of drug molecules. The DrugBank version 5.1.5 was used in this study which contains more than 13,000 drug entries, including approved small molecule drugs, biologics (proteins, peptides, vaccines, and allergenics), nutraceuticals and experimental (discovery-phase) drugs [(Wishart et al. 2018)](https://paperpile.com/c/a3wH5Q/loEpU). BLAST+ search was used to find the homologs of chokepoint under four target categories namely, drug targets, drug enzymes, drug carriers, and drug transporters. A drug target protein of DrugBank was considered as a homolog of chokepoint if it showed ≥50% identity over 80% of the sequence length. The drug molecules which showed the best hit of DrugBank target proteins were selected as potential binders of chokepoint proteins.

Further, drug molecules were optimized according to Lipinski's Rule of Five scales. The properties required by this rule are: molecular weight ≤500, number of rotatable bonds ≤10, hydrogen bond donors ≤5, hydrogen bond acceptors ≤10 and logP ≤5. In addition to Lipinski's rule of five, half-life ≥60 minutes and toxicity information of potential drug(s) were also considered. Only those drugs which qualified at least five of the seven parameters were included while dietary supplements, micronutrients, and vitamins were excluded from the list of potential drug molecules.

## 6.2.7. Analysis of drug-target interactions

To assess the binding potential of selected drug candidates with their target, PatchDock was used. PatchDock is a molecular docking algorithm that gives geometry shape complementarity score, area, atomic contact energy (ACE), and 3D transformation outputs [(Schneidman-Duhovny et al. 2005)](https://paperpile.com/c/a3wH5Q/1KSBE). The 3D structures of the MAP chokepoint protein were not found in PDB hence they were modeled using Swiss-model. The quality assessment of the homology models were done using several structure assessment parameters namely, Qualitative Model Energy Analysis (QMEAN) score [(Benkert et al. 2008)](https://paperpile.com/c/a3wH5Q/S54eo), MolProbity score [(Chen et al. 2010)](https://paperpile.com/c/a3wH5Q/WDbW4), and Ramachandran plot [(Ho & Brasseur 2005)](https://paperpile.com/c/a3wH5Q/tkTEv). Further, we also calculated the root-mean-square deviation (RMSD) between the modeled and the template structure.The structures of the potential drugs were downloaded from DrugBank. The protein-ligand complex with the highest docked score was selected and Ligplot+ was used for the analysis of binding interactions [(Wallace et al. 1995)](https://paperpile.com/c/a3wH5Q/MwPJw).

SiteEngine was used to recognize the functional binding sites of the protein models and compare it with the similar ligand binding pocket. SiteEngine maps the 4 **Å** region around the ligand and uses it to search for similar structural patterns on the surface of other proteins [(Shulman-Peleg et al. 2005)](https://paperpile.com/c/a3wH5Q/ZCOko). In the present work we used SiteEngine to compare the constituent amino acids of ligand-binding sites of modeled and experimentally determined structure of proteins. The information of proteins, whose structure was solved and are known to bind the same ligand, was obtained from the DrugBank.

**6.3. Results and Discussion**

## 6.3.1. Identification of interacting protein partners of MAP mimicry proteins

The keyword search “*Mycobacterium avium* subsp. *paratuberculosis*” in miPepBase displayed 14 entries/events related to mimicry. In these 14 events, 10 distinct proteins of MAP were found which exhibited molecular mimicry with the host proteins. These proteins cross-reacted with 10 different types of host proteins resulting in 5 different types of autoimmune diseases namely, encephalomyelitis, multiple sclerosis, autoimmune thyroiditis, Crohn’s disease and type 1 diabetes mellitus (Table E6.1). We found interacting protein partners for only eight of these 10 proteins in the STRING database (Table E6.2). The STRING database failed to provide information about interacting proteins of two MAP mimicry proteins (UniProt accession number: Q53467 and Q73SP6), hence these two proteins were removed from further analysis.

STRING analysis revealed that eight MAP mimicry proteins had 54 interacting protein partners (IPMMP), at a high confidence score. The individual interaction score and the final confidence score of each PPI are in Table E6.3. Of these, 16 IPMMPs were found to be homologous to human proteins hence, they were excluded from further analysis (Table 6.1).

**Table 6.1.** The list of MAP mimicry proteins, interacting partners of MAP mimicry proteins (IPMMP), name of human homolog of IPMMP (if present), Non-Human Homolog of IPMMP (nHIPMMP), KEGG pathway ID to which non-human homolog of IPMMP belongs (KEGG id is present in parenthesis), and chokepoint proteins.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S. No.** | **MAP mimicry proteins** | **IPMMP** | **HIPMMP** | **nHIPMMP** | **Pathway in which nHIPMMPs are involved (KEGG ID)** | **Chokepoint protein(s)** |
| 1 | P42384 | clpP2, dnaK, groEL1, groEL2, groES, grpE, guaA, gyrB, htpG, MAP\_2278c, rpoB | ClpP2, dnaK, groEL1, groEL2, groES, guaA, htpG, MAP\_2278c | grpE, gyrB, rpoB | RNA polymerase (mpa03020) | rpoB |
| 2 | Q73SP6 | NA | NA | NA | NA | NA |
| 3 | Q73T54 | MAP\_3865c, MAP\_3866c, MAP\_3867c | MAP\_3865c | MAP\_3866c, MAP\_3867c | NA | NA |
| 4 | Q73WG6 | fum, MAP\_2692, MAP\_2694 | fum | MAP\_2692, MAP\_2694 | Glycolysis / Gluconeogenesis (mpa00010)  Methane metabolism (mpa00680)  Pentose phosphate pathway (mpa00030)  Fructose and mannose metabolism (mpa00051) | NA |
| 5 | Q73WP1 | MAP\_0807c, MAP\_2102c, MAP\_2722c, MAP\_3707c, narG, narH, narI, narJ, narU, nirB, nirD | MAP\_0807c, | MAP\_2102c, MAP\_2722c, MAP\_3707c, narG, narH, narI, narJ, narU, nirB, nirD | Nitrogen metabolism (mpa00910)  Two-component system (mpa02020) | MAP\_2102c, MAP\_3707c, narG, narH, narI, narU, nirB, nirD, nirJ |
| 6 | Q73ZL3 | ahpC, ahpD, ahpE, catB, katG, oxyR, sodA, sodC, sucB, tpx, trxB2 | ahpC, ahpE, sodA, sucB | ahpD, catB, katG, oxyR, sodC, tpx, trxB2 | Selenocompound metabolism (mpa00450 )  Phenylalanine metabolism (mpa00360)  Drug metabolism - other enzymes (mpa00983)  Glyoxylate and dicarboxylate metabolism  (mpa00630)  Tryptophan metabolism (mpa00380) | katG,catB |
| 7 | Q740V8 | MAP\_1234, MAP\_1235, MAP\_3252 | NA | MAP\_1234, MAP\_1235, MAP\_3252 | Lipopolysaccharide biosynthesis (mpa00540) | MAP\_3252 |
| 8. | Q741P6 | apt, MAP\_1042, MAP\_1045, relA, secD, secE, secF, secY, yidC | apt | MAP\_1042, MAP\_1045, relA, secD, secE, secF, secY, yidC | Quorum sensing (mpa02024)  Protein export (mpa03060)  Bacterial secretion system (mpa03070)  Purine metabolism (mpa00230) | MAP\_1042, MAP\_1045, secD, secE, secF, secY |
| 9 | Q745A5 | MAP\_0106c, MAP\_2148, MAP\_2752 | NA | MAP\_0106c, MAP\_2148, MAP\_2752 | NA | NA |

**Table 6.2**. Drug and their target details.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pathogen mimicry protein** | **Chokepoint proteins** | **Chokepoint proteins which were part of** | | **Potential drug molecule as per DrugBank against choke point protein** | **Drugs that qualified at least 5 of 7 drug like properties** |
| **Essential gene database** | **Core proteome** |
| P42384 | rpoB | rpoB | rpoB | rpoB: DB04788; DB08226; DB08266; DB00615; DB01045; DB01220; DB04934; DB11753 | DB00615 (Rifabutin), DB01220 (Rifaximin), DB08226 (Myxopyronin B), DB08266 (Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene]carbamate) |
| Q73SP6 | NA | NA | NA | NA | NA |
| Q73T54 | NA | NA | NA | NA | NA |
| Q73WG6 | NA | NA | NA | NA | NA |
| Q73WP1 | MAP\_2102c, MAP\_3707c, narG, narH, narI, narU, nirB, nirD, nirJ | NA | MAP\_2102c, MAP\_3707c, narG, narH, narI, narU, nirB, nirD, nirJ | narH: DB04464; DB07349 | DB07349 (1S)-2-{[{[(2S)-2,3-DIHYDROXYPROPYL]OXY}(HYDROXY)PHOSPHORYL]OXY}-1-[(PENTANOYLOXY)METHYL]ETHYL OCTANOATE |
| Q73ZL3 | KatG,catB | katG | katG, catB | katG: DB00609; DB00951; DB08638 | DB00609 (Ethionamide), DB00951 (Isoniazid), DB08638 (2-AMINO-3-(1-HYDROPEROXY-1H-INDOL-3-YL)PROPAN-1-OL) |
| Q740V8 | MAP\_3252 | NA | NA | NA | NA |
| Q741P6 | MAP\_1042, MAP\_1045, secD, secE, secF, secY | secD, secE, secF, secY | MAP\_1042, MAP\_1045, secD, secE, secF, secY | NA | NA |
| Q745A5 | NA | NA | NA | NA | NA |

## 6.3.2. Identification and validation of the metabolic chokepoints of MAP

The 38 non-human homologs of interacting partners of MAP mimicry proteins were mapped on 17 metabolic pathways of MAP (Table 6.1). The pathways were analyzed manually to find possible chokepoint reaction(s). We found that these 38 proteins were a part of the 19 chokepoints of the MAP metabolic pathways (Table 6.1 and Figure E6.1).

Among the 19 chokepoint proteins, it was observed that 18 proteins were part of the core mycobacterial proteome (Table E6.4) and, 6 of the 19 chokepoint proteins shared a close homology with *M. tuberculosis* essential genes as per the DEG database. Thus, total 18 chokepoint proteins identified by us in this study qualified one or the other parameter of essentiality and were included in further analysis (Table 6.2).

## 6.3.3. Drug molecules for metabolic chokepoints

DrugBank search revealed 13 potential drug molecules against 3 of the 18 chokepoint proteins *viz.* katG, rpoB, and narH (Table 6.2). After benchmarking the drug molecules on the seven drug-like parameters, we were finally left with eight probable drug candidates (Table E6.5). Of these eight probable drug candidates, we noted that four molecules, *viz.* DB00609, DB00951, DB00615, and DB01220 were FDA-approved drugs and the remaining four molecules, *viz.* DB08638, DB08226, DB08266 and DB07349 were experimentally validated drugs (Table 6.3). The details of metabolic chokepoints of MAP and their corresponding DrugBank molecules are as follows:

(a) **katG:** katG was discerned as the chokepoint protein of the metabolic processes associated with MAP mimicry protein, alkylohydroperoxidase C or ahpC (UniProt Id: Q73ZL3). The ahpC protein of MAP mimics the human cytoskeleton-associated protein 5 (colonic and hepatic tumor overexpressed gene protein) resulting in CD [(Polymeros et al. 2006)](https://paperpile.com/c/a3wH5Q/3tZqL). katG is a bifunctional enzyme that shows both catalase and peroxidase activity [(Singh et al. 2008)](https://paperpile.com/c/a3wH5Q/tIIXI). Several studies have reported that katG protects *M. tuberculosis* from toxic reactive oxygen species, shows peroxynitritase activity, and helps in survival within the host macrophages [(Ng et al. 2004, Sherman et al. 1996, Wengenack et al. 1999)](https://paperpile.com/c/a3wH5Q/eazSA+ipn2x+iqSsE). We observed MAP-associated katG can be targeted by three DrugBank molecules *viz.* DB00609 (Ethionamide), DB00951 (Isoniazid) and DB08638. Of these, DB00951 (Isoniazid) has proved useful in treatment of *M. tuberculosis* and *M. avium intracellulare* infections [(Society & Research Committee of the British Thoracic Society 2001)](https://paperpile.com/c/a3wH5Q/Lgeht). Ethionamide (DB00609) is an FDA-approved drug, which is used in combination with other antituberculosis drugs in second-line treatment of multidrug-resistant active tuberculosis. It has been shown to be effective against *M. bovis* and *M. smegmatis* also [(Rastogi et al. 1996)](https://paperpile.com/c/a3wH5Q/Rm3cU). DB08638 (1-hydroperoxy-L-tryptophan) is an experimentally validated drug molecule that was reported to bind with the katG protein of *Burkholderia pseudomallei* (https://www.drugbank.ca/drugs/DB08638).

(b) **rpoB:** rpoB was identified as the chokepoint protein of metabolic processes associated with MAP mimicry protein hsp65 (UniProt Id: P42384) and its interacting partners. The hsp65 of MAP shows mimicry with human glutamate decarboxylase 2 which results in T1DM [(Naser et al. 2013)](https://paperpile.com/c/a3wH5Q/r7OT4). The *rpoB* gene encodes the beta-subunit of bacterial RNA polymerase. It was found to be the drug target of four DrugBank molecules *viz.* DB00615 (Rifabutin), DB01220 (Rifaximin), DB08226 (Myxopyronin B) and DB08266. Rifabutin (DB00615) and Rifaximin (DB01220) are FDA-approved, rifamycin antibiotics. Rifabutin is used for treating *M. avium* complex infections, while Rifaximin is used to treat traveler’s diarrhea, irritable bowel syndrome, and hepatic encephalopathy [(Rothstein 2016)](https://paperpile.com/c/a3wH5Q/jAVNN). Myxopyronin B (DB08226) and DB08266 are experimental drugs, which were shown to target rpoB of *Thermus thermophilus* (https://www.drugbank.ca/drugs/DB08266).

(c) **narH:** The MAP mimicry protein narH (UniProt Id: Q73WP1) mimics the human myelin oligodendrocyte glycoprotein which results in MS [(Cossu et al. 2014)](https://paperpile.com/c/a3wH5Q/XC2FW). Here, we found DrugBank molecule DB07349 as a potential inhibitor of narH. The narH gene encodes the beta chain of the enzyme nitrate reductase which helps the bacteria during anaerobic growth. The nitrate reductase enzyme complex has been studied extensively in *Escherichia coli* where it helps in using nitrate as an electron acceptor during anaerobic growth [(Ceccaldi et al. 2015)](https://paperpile.com/c/a3wH5Q/GgePA). DrugBank search revealed that the drug molecule DB07349 was already known to effectively target the narH protein of *E. coli* (https://www.drugbank.ca/bio\_entities/BE0003816)*.* DB07349 belongs to the class phosphatidylglycerols and is an experimental drug.

**Table 6.3**. Details of drugs proposed against MAP associated autoimmunity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Autoimmune Disease** | **MAP mimicry protein** | **Metabolic**  **Chokepoint** | **Drug Name** | **DrugBank ID** | **Drug Group** |
| Crohn’s disease | Q73ZL3 (ahpC) | katG | Ethionamide | DB00609 | FDA approved |
| Isoniazid | DB00951 | FDA approved |
| 2-AMINO-3-(1-HYDROPEROXY-1H-INDOL-3-YL)PROPAN-1-OL | DB08638 | Experimentally validated |
| Multiple Sclerosis | Q73WP1 (narH) | narH | 1S)-2-{[{[(2S)-2,3-DIHYDROXYPROPYL]OXY}(HYDROXY)PHOSPHORYL]OXY}-1-[(PENTANOYLOXY)METHYL]ETHYL OCTANOATE | DB07349 | Experimentally validated |
| Type 1 diabetes mellitus | P42384 (hsp65) | rpoB | Rifabutin | DB00615 | FDA approved |
| Rifaximin | DB01220 | FDA approved |
| Myxopyronin B | DB08226 | Experimentally validated |
| Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene] carbamate | DB08266 | Experimentally validated |

**Table 6.4.** Interaction pattern between drug and target proteins.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Selected target** | **Drug molecule** | **PatchDock score**  **(Kcal/mol)** | **Nature of interaction** | **Amino acid on binding sites** |
| katG | DB00609  (Ethionamide) | -101.17 | Hydrophobic interaction | Leu387, Pro241, Ser324, Arg110, Lys283, Trp330, His279, |
| Polar H interaction | Nill |
| DB00951  (Isoniazid) | -139.56 | Hydrophobic interaction | Ile109, Gly278, Trp113, His279, Gly282, Gly106, Phe281, Try330 |
| Polar H interaction | Lys283 |
| DB08638  (2-AMINO-3-(1-HYDROPEROXY-1H-INDOL-3-YL)PROPAN-1-OL) | -114.46 | Hydrophobic interaction | Gly106, Gly278, His279, Trp113, Asp143, Arg110, Lys283, Phe281 |
| Polar H interaction | Ser324 |
| narH | DB07349  (1S)-2-{[{[(2S)-2,3-DIHYDROXYPROPYL]OXY}(HYDROXY)PHOSPHORYL]OXY}-1-[(PENTANOYLOXY)METHYL]ETHYL OCTANOATE) | -11.62 | Hydrophobic interaction | Asn187, Ala192, Trp66, Leu74, Asp69, Gly80 |
|  | Polar H interaction | Arg81, Arg73, Arg75 |
| rpoB | DB00615  (Rifabutin) | -489.49 | Hydrophobic interaction | Lys 298, Asn292, Glu391, Leu87, Arg276, Arg299, Pro89, Glu91, Glu86, Thr288, Asp92, Arg395, Try272, Glu279, Ser94, Thr282, Glu284, Ser285, Thr288 |
| Polar H interaction | Pro280 |
| DB01220  (Rifaximin) | -45.14 | Hydrophobic interaction | Arg276, Arg395, Leu293, Gly203, Arg384, Ala204, Asn381, Gln382, Glu423, Ser388, Arg392, Leu289 |
| Polar H interaction | Arg299, Arg389, Val385, |
| DB08226  (Myxopyronin B) | -25.07 | Hydrophobic interaction | Ile90, Leu289, Arg395, Glu391, Glu91, Pro89, Asp221, Gly203, Ala204, Trp205, Arg373, Arg175, Val174, Arg384 |
| Polar H interaction | Arg299, Arg276 |
| DB08266  (Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene] carbamate) | 19.09 | Hydrophobic interaction | Arg373, Glu377, Arg384, Ser201, Asn381, Val385, Arg395, Glu391, Leu289, Glu91, Thr288 |
| Polar H interaction | Ser388, Arg299, Arg276 |

**Table 6.5.** Quality assessment scores of *in silico* protein models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein** | **% residues in the allowed regions of Ramachandran Plot** | **QMEAN score** | **MolProbity score** |
| katG | 97.25 | -0.06 | 1.09 |
| rpoB | 94.61 | 0.64 | 1.14 |
| narH | 97.89 | -2.64 | 0.78 |

## 6.3.4. Quality assessment of the modeled metabolic chokepoint protein structures

To evaluate the quality of the modeled proteins, we evaluated several structural features of the modeled structures. Analysis of the Ramachandran plot showed more than 94% residues of the modeled protein structures were present in the favored regions of the Ramachandran plot (Table 6.4), which is more than the required 90% threshold for a good quality protein structure [(Pereira et al. 2019)](https://paperpile.com/c/a3wH5Q/liFzT). The QMEAN server (https://swissmodel.expasy.org/qmean) provides a comprehensive quality score (Z-score) that measures the ‘degree of nativeness’ of the structural features of a modeled protein structure. The score also provides the likelihood that a given model is of comparable quality to experimental structures [(Benkert et al. 2011)](https://paperpile.com/c/a3wH5Q/F0MzQ). The QMEAN-score of the modeled structure of katG, rpoB, and narH were -0.06, 0.65, and -2.64 respectively, comparable to that of high-resolution experimental structures (Table 6.4). The MolProbity score provides a single quantifiable measure to assess the quality of a biomolecular structure. The score is calculated by a comprehensive all-atom contact analysis to find the steric problems within the query biomolecule [(Davis et al. 2007)](https://paperpile.com/c/a3wH5Q/IdHfN). The low MolProbity scores also indicated the good quality of the modeled structure (Table 6.4). Together, the three assessment methods confirmed that the modeled structures were of good quality. Further, when the protein models were superimposed with the corresponding templates and the RMSD of two structures were calculated, we found a very low RMSD value (0.118 for katG, 0.187 for narH, and 0.051 for rpoB). The structure validation scores and superimposed structures of target-template are shown in Figure E6.2.

SiteEngine webserver that was used to validate the drug binding site revealed a high similarity score and low RMSD values between the drug binding site of modeled protein and the binding site of the same drug on already reported target PDB structures (Table E6.6). Also, the amino acids present at the drug target interfaces of the modeled and experimentally determined protein structures were similar. These observations suggest that the proposed drug molecules bind to the appropriate binding sites in the protein models of katG, narH, and rpoB.

## 6.3.5. Protein-ligand interactions between the MAP proteins and DrugBank molecules

Binding analysis of katG with the identified DrugBank molecules (Table 6.5) revealed a better binding affinity for DB00951 (Isoniazid) (-139.56 kcal/mol) than for DB08638 (-114.46 kcal/mol) or DB00609 (Ethionamide) (-101.17 kcal/mol). Also, the number of residues bound by hydrogen and hydrophobic bonding in the katG-DB00951/DB08638 protein-ligand complex was more than the katG-DB00609 complex. In the case of rpoB, DB00615 (Rifabutin) showed a significantly higher binding energy (-489.49 kcal/mol) and, more hydrogen and hydrophobic binding residues in the protein-ligand complex than the other three DrugBank molecules, DB01220 (rifaximin), DB08226 (myxopyroninB) and DB08266. In the case of narH, the only identified drug molecule DB07349 showed a good binding affinity of -11.62 kcal/mol and strong polar and hydrophobic interactions between the protein and ligand. The interactions between each drug and target are shown in Figure E6.3.

## 6.3.6. Prediction of off-target binding

The targets of the proposed drug molecules are MAP proteins. Although MAP proteins which have a human homolog were omitted from this study, yet the presence of structurally conserved binding sites at the surfaces of human proteins might lead to “off-target” binding of the proposed drugs. Thus to remove this possibility, we used the PatchSearch webserver that searches for the potential “off-target” binding sites in a set of user-supplied protein structures. Further, it also estimates the binding affinity [(Rey et al. 2019)](https://paperpile.com/c/a3wH5Q/0keoh).

The PatchSearch “off-target” search results showed that the DrugBank molecules DB00609, DB00951 and DB08638 targeted against katG had a very poor binding affinity for human proteins (Table E6.7). The binding affinity of katG and the three DrugBank molecules ranged from -1.854 to -0.002, -0.256 to 0 and -0.381 to 0 kcal/mol, respectively. Similarly, the proposed DrugBank molecule DB07349 (target narH) also showed a negligible affinity for human proteins (range -3.141 to -0.006 kcal/mol). The four drug molecules namely, DB00615, DB01220, DB08226, DB08266 (target rpoB) also showed a very low binding affinity for human proteins (-0.446 to 0, -1.935 to 0, -10.25 to 0 and 0 kcal/mol respectively. Interestingly, two human proteins (PDB ID: 6D0M and 2YGN) showed a higher affinity for Myxopyronin B (DB08226). But, due to a non-significant higher RMSD value at the binding site (>1.5) it has a very low chance of “off-target” binding.

To summarize, in this study we have performed a systematic *in silico* analysis of MAP mimicry proteins and their interacting partners to identify chokepoints of their respective metabolic processes. Drug molecules, which qualified several stringent parameters, were shortlisted as appropriate inhibitors of the chokepoints. Molecular interactions between the drug target(s) and drug molecule(s) were further confirmed by molecular docking and off-target binding potential of the proposed drugs. Finally, we were left with eight DrugBank molecules which might prove useful for treating three MAP-associated autoimmune diseases, namely T1DM, CD and MS. Interestingly, all the drug molecules identified in our study, except Rifabutin are novel drug molecules which have not been tested for treating MAP-associated T1DM, CD, and MS. Moreover, the drug molecules identified during our analysis are either FDA-approved drugs or experimental drugs with proven efficacy. Hence, these can be easily incorporated in clinical studies or tested *in vitro* for assessing their suitability in treating MAP-associated autoimmune diseases. Thus, instead of proposing new chemotherapeutics, our study helps in repurposing the known drugs for treating MAP induced autoimmune T1DM, CD, and MS. We would like to add if other relevant databases like ChEMBL [(Gaulton et al. 2012)](https://paperpile.com/c/a3wH5Q/eenA1), PubChem [(Kim et al. 2016)](https://paperpile.com/c/a3wH5Q/XVirO), ChemBank [(Seiler et al. 2008)](https://paperpile.com/c/a3wH5Q/188CV), and ZINC database [(Irwin & Shoichet 2005)](https://paperpile.com/c/a3wH5Q/gqQ8m) were also included in the study, we might have discerned more drug molecules against these or other MAP-associated autoimmune diseases.

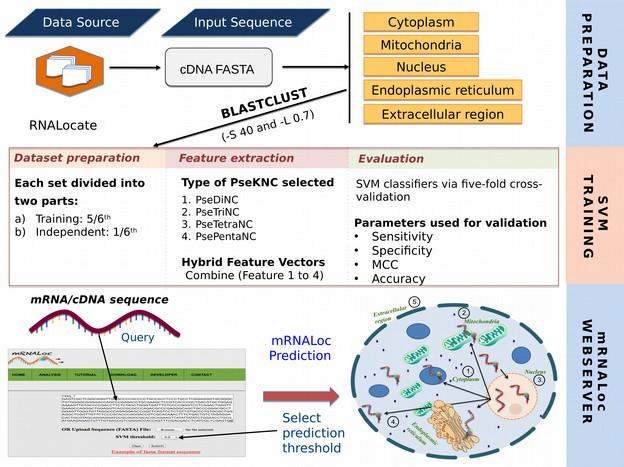
The results described in this chapter have been published as - Garg *et al*. Discerning novel drug targets for treating *Mycobacterium avium* ss. *paratuberculosis*-associated autoimmune disorders: an *in silico* approach. ***Briefings in Bioinformatics***, 2020.

# Chapter 7

# Additional work

**8.1. Introduction**

Localization of mRNA is an evolutionarily conserved phenomenon that controls many important biological processes like cell-fate determination and polar cell growth [(Kloc et al. 2002)](https://paperpile.com/c/a3wH5Q/7xmtf) . After post-transcriptional modifications, such as 5’ capping, splicing and addition of 3’ poly (A) tail, the nascent transcribed mRNA either gets localized within the nucleus or alternatively travels out of the nucleus. It has been suggested that mRNA localization has many advantages over protein localization [(Hughes & Simmonds 2019, Kaewsapsak et al. 2017, Lécuyer et al. 2007, Medioni et al. 2012)](https://paperpile.com/c/a3wH5Q/98Wcb+ntQJM+Tp6h6+0zS9I). These are: (a) localization of mRNA to a specific location helps the cell to build a local repository of proteins at the site of function instead of transporting individual protein molecules to the site of function. This also compartmentalizes protein synthesis and forms a protein gradient within the cells, which ultimately results in local synthesis of encoded proteins at the target site; (b) mRNA localization works as a translation/co-translational regulator; (c) mRNA localization is a better energy-efficient pathway compared to protein targeting and; (d) mRNA localization aids in formation of only functional and non-harmful multi-protein complexes which aids in avoiding unnecessary protein-protein interactions that might be harmful to the cells [(Liegro et al. 2014, Vandepoele et al. 2002)](https://paperpile.com/c/a3wH5Q/aVIcI+VwBzc) Not all protein synthesis occurs after mRNA localization. A large number of mRNA sequences are also transported co-translationally [(Weis et al. 2013)](https://paperpile.com/c/a3wH5Q/JSP4h). Five different mechanisms namely, diffusion and localized entrapment, localized degradation, localized synthesis, active transport and, polarized nuclear export are considered important for mRNA localization. However, ribonucleoprotein transport complex is the main mode by which majority of RNA is transported. Building the ribonucleoprotein complex is a sequence specific phenomenon, which is guided by a short stretch of 20-200 cis-acting nucleotide sequences known as ‘zipcode’. It is located at the 3′ untranslated region of the mRNA sequence, although in some cases they can also be present in the 5′UTR or in the coding sequence [(Heasman et al. 2001, Kloc & Etkin 1994)](https://paperpile.com/c/a3wH5Q/wxiLp+ranaK). Proteins present in a subcellular compartment are related to the physiological and metabolic function associated with that subcellular compartment. Hence, prediction of subcellular location of mRNA might suggests the biological function of the gene from which the mRNA was transcribed. Thus, a tool that can predict the correct intracellular location of transcripts may also help in understanding how gene expression is regulated and, how cells achieve polarity. To our knowledge, computational predictors that can predict the subcellular localization of eukaryotic mRNA are unavailable, till date. Hence, we describe a Support Vector Machine (SVM) based in-silico tool which can predict the eukaryotic mRNA subcellular locations on the basis of primary sequence information of mRNA/cDNA (Figure 8.1). Named as mRNALoc (acronym for “mRNA Localization”), this tool is based on the experimentally validated localization data of mRNA retrieved from ‘RNALocate’ [(Zhang et al. 2017)](https://paperpile.com/c/a3wH5Q/d7fbS).



**Figure 8.1.** Systematic workflow of mRNAloc training and functioning. (Adapted from: Garg *et al* (2020))

# 8.2. Material and methods

## 8.2.1. Collection of subcellular location annotated dataset of mRNA

Constructing a high quality benchmark dataset is the foremost requirement to develop a reliable prediction model. For the present work we collected the mRNA sequences and their subcellular location information from RNALocate database [(Zhang et al. 2017)](https://paperpile.com/c/a3wH5Q/d7fbS). Initially a total of 28829 mRNA sequences with annotated subcellular localization were obtained. This included multi-locations mRNA also. On the basis of subcellular locations, the mRNA were classified into five subgroups namely, cytoplasmic, endoplasmic reticulum, extracellular, mitochondrial and nuclear. The number of mRNA sequences in the five locations were as follows: 6964 in cytoplasm, 1998 in endoplasmic reticulum, 1131 in extracellular region, 442 in mitochondria and 6346 in nucleus.

## 8.2.2. Redundancy removal

The sequence redundancy among mRNA might results in overestimation of prediction capability. Hence to get rid of the redundancy and to avoid homology bias in prediction, we used NCBI BLASTCLUST program to keep only sequences having alignment identity ≤ 40% over 70% or more of their full length (BLASTCLUST with “-S 40 and -L 0.7” option) [(McGinnis & Madden 2004)](https://paperpile.com/c/a3wH5Q/TkVk7). The final non-redundant mRNA dataset contained 6376 sequences of cytoplasm, 1426 sequences of endoplasmic reticulum, 855 sequences of extracellular region, 421 sequences of mitochondria and 5831 sequences of nucleus. 5/6 part of total 40% non-redundant data was used for training the model. Remaining 1/6 data was used for the independent evaluation of the trained model. The NCBI gene accession numbers of mRNA and their sequences and locations are available at mRNALoc Webserver (<http://proteininformatics.org/mkumar/mrnaloc/download.html>).

## 8.2.3. Construction of training and independent dataset and training methodology

We divided redundancy reduced mRNA sequences of each subcellular location into two parts. The subset that contained 5/6 th of total data was used as training dataset for SVM while the remaining 1/6 th part was used as independent dataset to assess the performance of trained model. For training we adopted five-fold cross-validation approach during which all mRNA sequences of training dataset were randomly divided into five sets of which four sets were used for training and the remaining one set for testing. This procedure was repeated five times in such a way that each set was used once for testing.

## 8.2.4. Conversion of a nucleotide sequence into machine learning input feature

Since SVM can be trained with only fixed length input features, we converted each mRNA sequence to a fixed length numerical encodings. The most simple feature encoding is nucleotide compositions of an mRNA. However the simple nucleotide composition does not contain the sequence-order effect. To incorporate local sequence-order information K-tuple nucleotide composition (PseKNC) [(Chen et al. 2014)](https://paperpile.com/c/a3wH5Q/mtR5f) was proposed, which has been used in development of many biological feature predictors [(Chen et al. 2015, Liu et al. 2015, 2016, 2017a)](https://paperpile.com/c/a3wH5Q/sTjIY+0kLtC+gzmWz+WLwBU). The value and dimension of PseKNC depends on the value of K that ranges from 2, 3, 4 and 5 for di-, tri-, tetra- and penta-nucleotide respectively. It was observed that DNA/RNA dinucleotide physical structures, including twist, tilt, roll, shift, slide and rise, significantly contribute to dealing with DNA/RNA sequences [(Liu et al. 2017b, 2018)](https://paperpile.com/c/a3wH5Q/CR7z0+vW8bf). Therefore, in this study six dinucleotide physical structures are employed to encode the pseudo 2-tuple nucleotide compositions. Additionally, 12 physicochemical properties of tri-, tetra- and pentanucleotide were also included to encode the pseudo 3, 4, 5-tuple nucleotide compositions.

## 8.2.5. Hybrid Feature Vectors (HFV)

Several studies have shown that a SVM model that is trained on a combination of more than one input features has better discrimination capability [(Kumar & Raghava 2009, Kumar et al. 2006, Mishra et al. 2007, Tahir et al. 2018)](https://paperpile.com/c/a3wH5Q/ilyca+4SAH9+nr4vS+y2ntf). Hence we combined and used all input features (namely PseDNC, PseTNC, PseTetraNC, and PsePentaNC) as a single hybrid input feature vector. The combined hybrid feature vector had the size 1360-D (16+64+256+1024) for an mRNA sequence.

## 8.2.6. Fragmented Sequence Encoding

In our earlier work we have observed that when a protein sequence was divided into multiple fragments and a combined input vector was made from the sequence encoding of individual fragments, the performance of predictor increased significantly [(Kumar et al. 2006, 2014, 2018)](https://paperpile.com/c/a3wH5Q/4SAH9+Naxum+KSNgL). Hence in the present work we have also used the fragmented sequence approach to train SVM. We split each mRNA sequence into three equal parts and named them as N-terminal, M-part and C-terminal, each part individually used for feature identification. For example an mRNA sequence of length L with N number of nucleotides is divided into three chunks where first L/3 rd part is used as N-terminal and second L/3 rd part as M-part and last L/3 rd as C-terminal. For each fragment we calculated the values of PseKNC and used the combination of PseKNC of three segments as input to the SVM. We used both single and hybrid approach of encoding as an input to SVM.

## 8.2.7. Performance Evaluation Strategies

In this study five-fold cross-validation approach was used to test the performance of SVM models after training. For training mRNA of each location was randomly divided into five approximately equal-sized non-overlapping subsets. At each parameter the SVM was trained on four subsets and tested on the remaining one subset. This procedure was repeated five times so that every time a different subset was used as the test set and each subset were used as a test set only once. The performance at each parameter was measured for each test, and the average of five test subsets was reported as the final performance.

The classification performance for the subcellular location of mRNA was evaluated using sensitivity (Sn), specificity (Sp), Matthew’s correlation coefficient (MCC) and accuracy (ACC). All of these parameters were derived from the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These parameters are formulated as:

(1)

(2)

(3)

(4)

In eq. 1..4 for a particular location X, TP was the number of mRNA sequence correctly predicted to be present in X location, TN represent mRNA which were correctly predicted to be not present in X location, FP is the number of sequences belong to location Y but wrongly predicted as location X and FN is the mRNA sequences which actually belong to location X but predicted as Y location. To describe the performance of models across the entire range of SVM decision values, receiver operating characteristics (ROC) curve [(Sing et al. 2005)](https://paperpile.com/c/a3wH5Q/GPil0) and area under ROC curve (AUC) were used. ROC curves showed the true positive rate as a function of the false positive rate. The area under the ROC curve (the AUC score) is a way to transform the information provided by ROC curve to a single scalar value representing expected performance. Random prediction by a SVM model will have ROC curve at the diagonal line with AUC score of 0.5, while a perfect predictor will produce a ROC curve along the left and top boundary of the square and will have AUC score one.

# 8.3. Overview of mRNALoc

mRNALoc is a web resource to predict the subcellular localization of eukaryotic mRNA. The overall workflow of mRNALoc is shown in Figure 8.2. Users have to provide the mRNA sequences in a FASTA format. The submitted mRNA sequence will be converted into numerical encoding using pseudo oligonucleotide composition or pseudo K-tuple nucleotide composition (PseKNC) [(Chen et al. 2015, Liu et al. 2015, 2016, 2017a)](https://paperpile.com/c/a3wH5Q/WLwBU+sTjIY+0kLtC+gzmWz). On the basis of SVM prediction score, the mRNA will be predicted to localize at one of the five subcellular locations, namely cytoplasm, endoplasmic reticulum, extracellular location, mitochondria and nucleus. mRNALoc prediction is based on the five trained SVM models, each specific for one location. During prediction each model provides the prediction score for its corresponding location. The subcellular location, whose SVM model gets the maximum score, will be the predicted location. The final outcome of mRNALoc depends on the user-selected threshold. Higher thresholds would result in more specific predictions, while lower threshold would result in low specificity predictions.

# 8.4. Results and Discussion

## 8.4.1. Performance on Complete Sequence Information

In all five locations firstly we used PseDNC (K=2) as input. Subsequently the value of K increased to five. We noticed a significant improvement in performance when value of K increased from 2 to 4 and after K=4 no increase in performance was observed. Performance on Fragmented Sequence Information. Similar to the complete sequence information we observed a considerable improvement in performance when value of K increased from 2 to 4 and after K=4 no increase in performance was observed. However we didn’t notice any improvement in performance when fragmented PseKNC was used as input instead of complete sequence information.

## 8.4.2. All Features Helped, but the Combination Performed Best and Most Robust

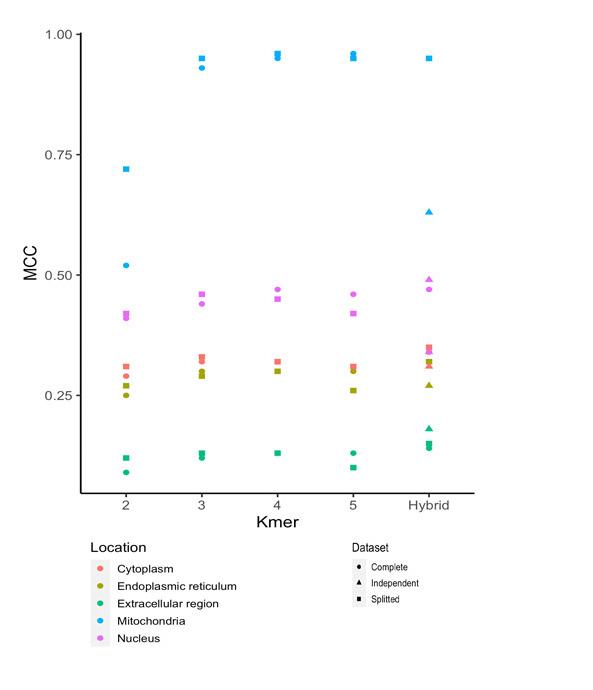
When all five forms of PseKNC were merged together as a single input, we noticed a significant improvement in performance in case of both complete and fragmented sequence input. In most locations an increase of 2% was observed with hybrid inputs both in case of complete and splitted PseKNC (Figure 8.3).

## 8.4.3. Receiver Operating Characteristics (ROC) Plot and Area Under ROC Curve (AUC) analysis

Use of overall accuracy to estimate the performance of a predictor developed on an imbalanced dataset might provide an unrealistic assessment of a classifier’s performance because even a random one-sided prediction of majority class members may create an impression of a highly accurate predictor, which is infact based on an one-sided random predictor. In the present work, to avoid the influence of majority data types on prediction efficiency calculations, we gave equal weightage to both sensitivity and specificity. Another way of objective estimation of classifier accuracy is receiver operating characteristic (ROC) plot, which is a very popular way to measure overall performance of a classifier. ROC plot demonstrates the trade-off between sensitivity and specificity at different thresholds and is generated by plotting ‘True positive rate’ vs. ‘False positive rate’. Further, the area under the ROC curve (AUC) is also frequently used to estimate the performance of a classifier. As shown in figures 8.4 and table 8.1, the ROC plot and their corresponding AUC values also supports the conclusion that SVM model developed using hybrid composition as input has performed better in comparison to the ones which were developed using a single input. Also since performance of both complete sequence and splitted composition based SVM modules are same, in all further studies we used hybrid full sequence composition based SVM modules.

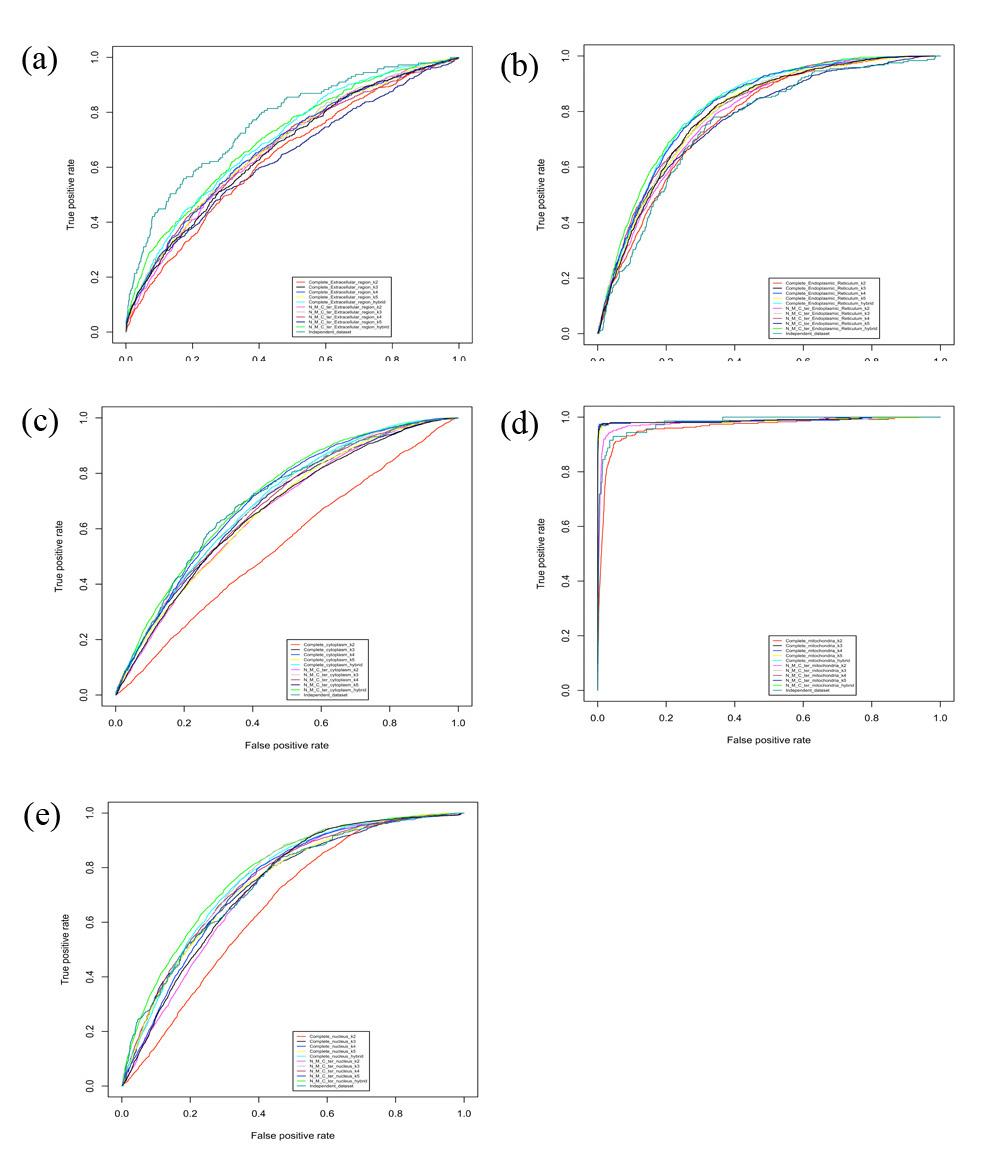


**Figure 8.2. Overall schema of mRNALoc.** mRNALoc predicts five subcellular locations viz., mitochondria, cytoplasm, nucleus, endoplasmic reticulum and extracellular. Initially it removes non-standard nucleotides from the sequence, generates combined features from pseudo K-tuple nucleotide composition which is further used as input for Support Vector Machine (SVM) prediction. (Adapted from: Garg *et al* (2020))

**Figure 8.3.** Performance achieved during five-fold cross-validation with different values of K-mer. (Adapted from: Garg *et al* (2020))

**Table 8.1.** Estimation of the performance metrics for mRNA location identification under different combination of K-mer features. Values indicated in italics and bold font show the highest prediction score of training, and only in bold font show the performance on independent dataset. (Sen: Sensitivity, Spe: Specificity, ACC: Accuracy, MCC: Mathew’s correlation coefficient, TN: True negative, FN: False negative, TP: True positive, FP: False positive, THR: Threshold, and AUC: Area under ROC curve)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Complete sequence** | **Sen (%)** | **Spe (%)** | **ACC**  **(%)** | **MCC** | **TN** | **FN** | **TP** | **FP** | **THR** | **AUC** | **K-mer** | **# of features** | **Location** |
| 58.17 | 61.08 | 60.91 | 0.09 | 7146 | 297 | 413 | 4554 | -0.2 | 0.63 | 2 | 16 | **Extracellular region** |
| 60.00 | 63.93 | 63.70 | 0.12 | 7479 | 284 | 426 | 4221 | -0.4 | 0.66 | 3 | 64 |
| 61.27 | 65.27 | 65.04 | 0.13 | 7636 | 275 | 435 | 4064 | -0.4 | 0.67 | 4 | 256 |
| 62.96 | 63.36 | 63.34 | 0.13 | 7413 | 263 | 447 | 4287 | -0.6 | 0.67 | 5 | 1024 |
| ***62.67*** | ***65.34*** | ***65.19*** | ***0.14*** | ***7645*** | ***265*** | ***445*** | ***4055*** | ***-0.2*** | ***0.69*** | ***2+3+4+5*** | 1360 |
| 70.30 | 70.41 | 70.40 | 0.25 | 7904 | 352 | 833 | 3321 | 0.4 | 0.77 | 2 | 16 | **Endoplasmic reticulum** |
| 72.49 | 73.53 | 73.43 | 0.30 | 8254 | 326 | 859 | 2971 | 0.1 | 0.79 | 3 | 64 |
| 74.26 | 74.86 | 74.80 | 0.32 | 8403 | 305 | 880 | 2822 | 0.2 | 0.81 | 4 | 256 |
| 72.40 | 73.68 | 73.56 | 0.30 | 8271 | 327 | 858 | 2954 | 0 | 0.79 | 5 | 1024 |
| ***74.09*** | ***75.49*** | ***75.36*** | ***0.32*** | ***8474*** | ***307*** | ***878*** | ***2751*** | ***0.4*** | ***0.81*** | ***2+3+4+5*** | 1360 |
| 62.26 | 66.53 | 64.71 | 0.29 | 4724 | 2004 | 3306 | 2376 | 0.5 | 0.54 | 2 | 16 | **Cytoplasm** |
| 66.05 | 66.35 | 66.22 | 0.32 | 4711 | 1803 | 3507 | 2389 | 0.4 | 0.66 | 3 | 64 |
| 65.37 | 66.80 | 66.19 | 0.32 | 4743 | 1839 | 3471 | 2357 | -0.4 | 0.70 | 4 | 256 |
| 65.56 | 65.38 | 65.46 | 0.31 | 4642 | 1829 | 3481 | 2458 | 0.3 | 0.66 | 5 | 1024 |
| ***66.69*** | ***67.41*** | ***67.10*** | ***0.34*** | ***4786*** | ***1769*** | ***3541*** | ***2314*** | ***0.4*** | ***0.69*** | ***2+3+4+5*** | 1360 |
| 91.43 | 93.68 | 93.62 | 0.52 | 11298 | 30 | 320 | 762 | -0.3 | 0.96 | 2 | 16 | **Mitochondria** |
| 96.00 | 99.69 | 99.59 | 0.93 | 12023 | 14 | 336 | 37 | -0.1 | 0.98 | 3 | 64 |
| 97.14 | 99.77 | 99.70 | 0.95 | 12032 | 10 | 340 | 28 | -0.3 | 0.98 | 4 | 256 |
| 95.14 | 99.89 | 99.75 | 0.96 | 12046 | 17 | 333 | 14 | -0.2 | 0.98 | 5 | 1024 |
| ***96.28*** | ***99.79*** | ***99.70*** | ***0.95*** | ***12035*** | ***13*** | ***337*** | ***25*** | ***0.1*** | ***0.98*** | ***2+3+4+5*** | 1360 |
| 69.56 | 71.83 | 70.94 | 0.41 | 5427 | 1478 | 3377 | 2128 | -0.3 | 0.66 | 2 | 16 | **Nucleus** |
| 72.62 | 72.49 | 72.55 | 0.44 | 5477 | 1329 | 3526 | 2078 | 0.4 | 0.73 | 3 | 64 |
| 75.08 | 72.27 | 73.37 | 0.47 | 5460 | 1210 | 3645 | 2095 | 0.3 | 0.74 | 4 | 256 |
| 74.07 | 72.35 | 73.02 | 0.46 | 5466 | 1259 | 3596 | 2089 | -0.3 | 0.74 | 5 | 1024 |
| ***74.17*** | ***73.22*** | ***73.59*** | ***0.47*** | ***5532*** | ***1254*** | ***3601*** | ***2023*** | ***0.4*** | ***0.76*** | ***2+3+4+5*** | 1360 |
| **Splitted sequence** | 60.42 | 63.90 | 63.70 | 0.12 | 7476 | 281 | 429 | 4224 | -0.1 | 0.66 | 2 | 48 | **Extracellular region** |
| 62.39 | 64.31 | 64.21 | 0.13 | 7525 | 267 | 443 | 4175 | -0.3 | 0.67 | 3 | 192 |
| 61.69 | 64.09 | 63.95 | 0.13 | 7498 | 272 | 438 | 4202 | -0.5 | 0.67 | 4 | 768 |
| 59.01 | 61.86 | 61.70 | 0.10 | 7238 | 291 | 419 | 4462 | -1 | 0.63 | 5 | 3072 |
| ***64.37*** | ***65.42*** | ***65.36*** | ***0.15*** | ***7654*** | ***253*** | ***457*** | ***4046*** | ***-0.3*** | ***0.70*** | ***2+3+4+5*** | ***4080*** |
| 71.22 | 71.19 | 71.19 | 0.27 | 7991 | 341 | 844 | 3234 | -0.1 | 0.78 | 2 | 48 | **Endoplasmic reticulum** |
| 72.15 | 72.93 | 72.85 | 0.29 | 8186 | 330 | 855 | 3039 | 0.2 | 0.79 | 3 | 192 |
| 73.58 | 73.59 | 73.59 | 0.30 | 8261 | 313 | 872 | 2964 | -0.1 | 0.80 | 4 | 768 |
| 69.11 | 71.87 | 71.60 | 0.26 | 8067 | 366 | 819 | 3158 | -0.3 | 0.76 | 5 | 3072 |
| ***73.16*** | ***75.42*** | ***75.21*** | ***0.32*** | ***8466*** | ***318*** | ***867*** | ***2759*** | ***0.3*** | ***0.82*** | ***2+3+4+5*** | ***4080*** |
| 67.82 | 63.90 | 65.58 | 0.31 | 4537 | 1709 | 3601 | 2563 | 0.4 | 0.66 | 2 | 48 | **Cytoplasm** |
| 66.93 | 66.30 | 66.57 | 0.33 | 4707 | 1756 | 3554 | 2393 | 0.4 | 0.69 | 3 | 192 |
| 66.67 | 65.65 | 66.08 | 0.32 | 4661 | 1770 | 3540 | 2439 | -0.4 | 0.68 | 4 | 768 |
| 66.86 | 63.96 | 65.20 | 0.31 | 4541 | 1760 | 3550 | 2559 | 0.2 | 0.68 | 5 | 3072 |
| ***67.02*** | ***68.38*** | ***67.80*** | ***0.35*** | ***4855*** | ***1751*** | ***3559*** | ***2245*** | ***0.4*** | ***0.71*** | ***2+3+4+5*** | ***4080*** |
| 93.14 | 97.82 | 97.69 | 0.72 | 11797 | 24 | 326 | 263 | -0.5 | 0.98 | 2 | 48 | **Mitochondria** |
| 96.00 | 99.81 | 99.70 | 0.95 | 12037 | 14 | 336 | 23 | -0.1 | 0.98 | 3 | 192 |
| 96.00 | 99.88 | 99.78 | 0.96 | 12046 | 14 | 336 | 14 | -0.1 | 0.98 | 4 | 768 |
| 95.71 | 99.86 | 99.74 | 0.95 | 12043 | 15 | 335 | 17 | -0.3 | 0.98 | 5 | 3072 |
| ***95.43*** | ***99.82*** | ***99.70*** | ***0.95*** | ***12038*** | ***16*** | ***334*** | ***22*** | ***0.1*** | ***0.98*** | ***2+3+4+5*** | ***4080*** |
| 72.48 | 70.81 | 71.47 | 0.42 | 5350 | 1336 | 3519 | 2205 | -0.3 | 0.72 | 2 | 48 | **Nucleus** |
| 74.17 | 72.31 | 73.04 | 0.46 | 5463 | 1254 | 3601 | 2092 | -0.3 | 0.76 | 3 | 192 |
| 71.70 | 73.88 | 73.03 | 0.45 | 5582 | 1374 | 3481 | 1973 | 0.3 | 0.75 | 4 | 768 |
| 71.45 | 70.79 | 71.05 | 0.42 | 5348 | 1386 | 3469 | 2207 | -0.3 | 0.74 | 5 | 3072 |
| ***75.26*** | ***74.19*** | ***74.61*** | ***0.49*** | ***5605*** | ***1201*** | ***3654*** | ***1950*** | ***0.3*** | ***0.78*** | ***2+3+4+5*** | ***4080*** |
| **Independent dataset** | **81.38** | **56.67** | **58.10** | **0.18** | **1334** | **27** | **118** | **1020** | **-0.2** | **0.76** | **2+3+4+5** | **1360** | **Extracellular region** |
| **75.10** | **68.6** | **69.23** | **0.27** | **1549** | **60** | **181** | **709** | **0.4** | **0.75** | **2+3+4+5** | **1360** | **Endoplasmic reticulum** |
| **73.26** | **58.06** | **64.55** | **0.31** | **832** | **285** | **781** | **601** | **0.4** | **0.70** | **2+3+4+5** | **1360** | **Cytoplasm** |
| **87.32** | **97.16** | **96.88** | **0.63** | **2359** | **9** | **62** | **69** | **0.1** | **0.98** | **2+3+4+5** | **1360** | **Mitochondria** |
| **50.20** | **81.62** | **69.35** | **0.34** | **1243** | **486** | **490** | **280** | **0.4** | **0.74** | **2+3+4+5** | **1360** | **Nucleus** |



**Figure 8.4.** The ROC curves for (a) extracellular region (b) endoplasmic reticulum (c) cytoplasm (d) mitochondria (e) nucleus. A ROC cure plots the true positive rate (i.e. sensitivity) against the false positive rate. For each location performance on complete and splitted sequence is shown at different Kmer values. The performance on independent data is also shown. (Adapted from: Garg *et al* (2020))

## 8.4.4. Benchmarks on independent mRNA datasets

We also evaluated the performance of mRNALoc hybrid full sequence composition based SVM models on an independent dataset. The data was 1/6 th part of the original data collected initially and it was not was used to train the SVM. On same prediction parameters, it gave prediction result with sensitivity, specificity, accuracy and MCC values 75.10, 68.60, 69.23 and 0.27 for endoplasmic reticulum, 81.38, 56.67, 58.1 and 0.18 for extracellular region, 73.26, 58.06, 64.55 and 0.31 for cytoplasm, 87.32, 97.16, 96.88 and 0.63 for mitochondria and 50.20, 81.62, 69.35 and 0.34 for nucleus respectively.

# 8.5. Comparison with existing mRNA subcellular localization prediction methods

Though, the role of mRNA localization is unambiguously established in cellular physiology, attempts to build in-silico tools to predict the subcellular localizations of mRNA are negligible in comparison to protein subcellular localization prediction tools. Recently, Yan et al proposed a deep-learning based method, named as RNATracker [(Yan et al. 2019)](https://paperpile.com/c/a3wH5Q/y0tdN), to predict the subcellular localization of mRNA using data from CeFra-Seq [(Bouvrette et al. 2018)](https://paperpile.com/c/a3wH5Q/PkFTy) and APEX-RIP [(Kaewsapsak et al. 2017)](https://paperpile.com/c/a3wH5Q/98Wcb) Using the data from RNALocate, a human mRNA subcellular localization method iLoc-mRNA was developed [(Bouvrette et al. 2018)](https://paperpile.com/c/a3wH5Q/PkFTy) Though, both RNATracker and iLoc-mRNA are based on two different mRNA subcellular localization datasets and, were developed using two different approaches, mRNALoc has several advantages over both RNATracker and iLoc-mRNA. For example, (a) localization data produced by CeFra-Seq/APEX-RIP are inherently noisy and sometimes inaccurate, also [(Yan et al. 2019)](https://paperpile.com/c/a3wH5Q/y0tdN). The mRNALoc was developed from datasets retrieved from RNALocate [(Zhang et al. 2017)](https://paperpile.com/c/a3wH5Q/d7fbS) which contains manually curated mRNA subcellular localization information with experimental evidences. (b) The RNATracker among all the isoforms, considered only the longest isoform while, mRNALoc did not made any such distinction. (c) Redundant mRNA sequences were not removed from RNATracker and in iLoc-mRNA the redundancy threshold was 80%. While in mRNALocate, we used 40% non-redundant mRNA sequences to train the predictor. This may be the reason underlying high MCC and AUC for RNATracker and iLoc-mRNA. (d) Both, RNATracker and iLoc-mRNA were developed using only localization data of human mRNA. On the contrary, mRNALoc is a general-purpose eukaryotic mNRA subcellular localization prediction tool, which is applicable to all eukaryotes. (e) RNATracker also excluded low expressed genes, but mRNALoc made no such distinction (Table 8.2).

We also conducted one-to-one comparison of performance of iLoc-mRNA and mRNALoc As RNATracker required gene expression and coordination files for prediction, it was not possible to include it in the evaluation. For comparison we used the independent dataset of mRNALoc. Since, iLoc-mRNA is specifically designed for human mRNA subcellular localization prediction, we used 50 human mRNA sequences of independent dataset of mRNALoc. The number of human mRNA in different locations and prediction result of mRNALoc and iLoc-mRNA is shown in Table 8.3. In extracellular region and mitochondria, we didn’t find human mRNA sequences in mRNALoc independent dataset hence, these locations were not included in the evaluation.

As shown in table 3, for cytoplasm and nucleus the performance of mRNALoc was better than iLoc-mRNA but, in endoplasmic reticulum the performance of iLoc-mRNA was better than mRNALoc. It is also pertinent to mention that in iLoc-mRNA prediction were made for one of the following locations namely, cytosol/cytoplasm, ribosome, endoplasmic reticulum, and nucleus/exosome/dendrite/mitochondrion. We feel that combining nucleus, exosome, dendrite, and mitochondria as a single location is not appropriate as these are diverse subcellular locations

which should not be merged in a single category.

**Table 8.2.** Brief comparison of the advantages and limitations among mRNALoc, RNATracker and iLoc-mRNA.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **mRNALoc** | **RNATracker** | **iLoc-mRNA** |
| *Benchmark data source* | RNALocate | CeFra-Seq/APEX-RIP | RNALocate |
| *Data redundancy threshold* | 40% | 80% | 80% |
| Types of mRNA sequences incorporated in the study | All types of genes and isoforms | Only highly expressed and longest isoforms | Not mentioned |
| *Tool used to develop prediction Model* | SVM | Deep neural network | SVM |
| *Can work on* | All Eukaryotes | Human only | Human only |
| *Input data* | mRNA/cDNA | Sequencing read and transcript coordination file | mRNA/cDNA |
| *Webserver* | Yes | No | Yes |
| *Standalone availability* | Yes | Yes | Yes |

**Table 8.3.** Comparative evaluation of mRNALoc and iLoc-mRNA. In extracellular region and mitochondria no human mRNA was present, hence these two locations were not included in the evaluation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Location | Number of human mRNA sequences | mRNALoc | | iLoc-mRNA | |
| Correct | Wrong | Correct | Wrong |
| Cytoplasm | 50 | 35 | 15 | 18 | 32 |
| Endoplasmic reticulum | 50 | 34 | 16 | 37 | 13 |
| Extracellular region | 0 | 0 | 0 | 0 | 0 |
| Mitochondria | 0 | 0 | 0 | 0 | 0 |
| Nucleus | 50 | 33 | 17 | 13 | 37 |

# 8.6. Description of the webserver

## 8.6.1. Implementation of mRNALoc

The web server is hosted on a Linux system. The back-end pipeline is implemented in the Perl language. The webserver has an intuitive interface and ‘how-to’ guide to help the user. Each mRNA query sequence must be at least 100 bp long and contains only valid characters, namely ‘A’, ‘C’, ‘G’ and ‘T/U’. Sequences having non-standard nucleotides will be omitted from the prediction pipeline (Figure 8.5).

## 8.6.2. The output of mRNALoc

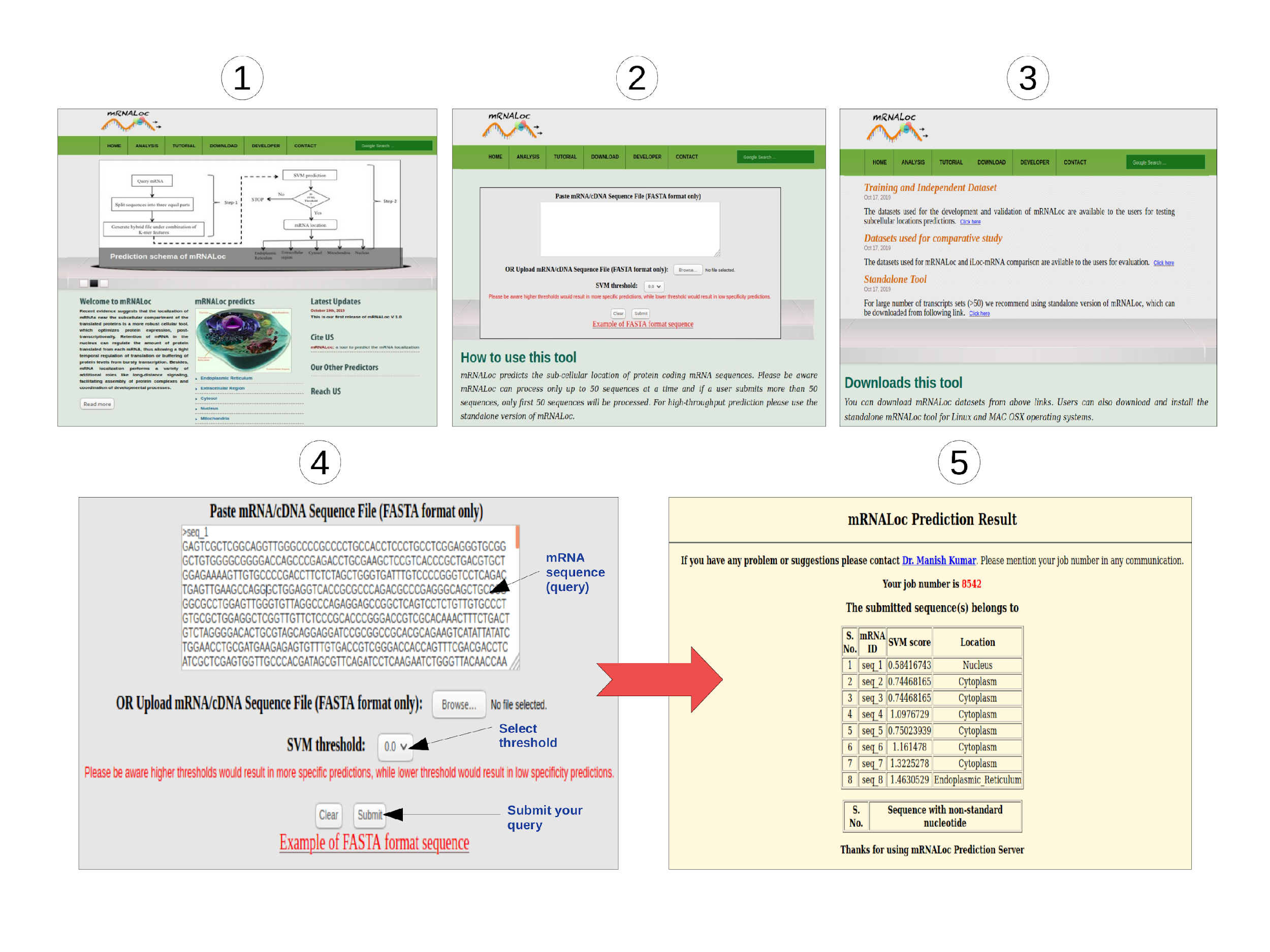
The output of mRNALoc is presented in a tabular format. It contains the highest scores obtained

from the five SVM models and the location to which the mRNA is assigned. A maximum of fifty

sequences can be processed by mRNALoc webserver in one go. Hence, for genome scale prediction a standalone version will be required (Figure 8.5).

**8.7. Conclusions and future prospects**

The annotation of subcellular localization has been addressed mainly at the protein level. Many *in silico* tools were developed to predict protein subcellular location using machine-learning techniques. It has been unequivocally established that both mRNA and protein localization play an equal role in protein translocation. In future versions of mRNALoc we would like to overcome some of the limitations of the present tool. The first and foremost is that our tool is currently limited by the accuracy of the RNALocate datasets. Though, RNALocate contain data from 65 organisms, most of the data is enriched with the common model organisms like, *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae* etc. Moreover, considering at the biological level, instead of cytosol, mitochondria or extracellular locations, axons, dendrites, dendritic spines, or anterior/posterior vs dorsal/ventral locations are more relevant. Another, limitation is that due to lesser availability of plant mRNA localization data compared to other domains of life, mRNALoc performance might be compromised [(Tian et al. 2020)](https://paperpile.com/c/a3wH5Q/FlMjU). The performance of a machine-learning method depends on the data on which it is trained. We believe that with development of new and better RNA localization finding techniques, information about RNA localization in plants would also be available in the near future and future versions of mRNALoc would then support prediction of plant mRNA sequences, also. We admit that mRNALoc is in an early stage of development and training on additional datasets is needed to further improve our tool. Further prediction of mRNA localization will also help in predicting the novel zipcodes which may guide researchers to cast new hypothesis for unraveling the finer details of mechanism of mRNA-protein complex formation which is actually responsible for mRNA location. Though, the current version of mRNALoc supports prediction of only eukaryotic mRNA, the future versions of mRNALoc would definitely include data from other organisms and locations.



**Figure 8.5. Screenshots of mRNALoc webserver.** (Adapted from: Garg *et al*. (2020))

The results described in this chapter have been published as - Garg *et al*. mRNALoc: a novel machine-learning based *in-silico* tool to predict mRNA subcellular localization. ***Nucleic Acids Research***, 2020.

# Summary & Future prospect

Mimicry is the resemblance of one life form with another life form, and it provides a selective advantage to the mimicker. Mimicry can involve physical or behavioural traits, and well-studied examples are Batesian and Mullerian mimicry. Mimicry as observed in nature is not only an ecological phenomenon, it is observed in the microbial world too!! Molecular mimicry can be defined as structural, functional, or immunological similarity between the host and pathogen macromolecules. Molecular mimicry can be present in the form of complete identity or homology at the protein level, or as a similarity in the sequences of amino acids or as structural similarity of host and pathogen proteins. Sometimes the host develops an immune response against self-proteins involved in molecular mimicry, which results in autoimmune diseases.

The work described in this thesis explains studies related to sequential molecular mimicry based autoimmune diseases. As we know some autoimmune diseases affect the rate of morbidity and mortality in the human population. The studies presented in this thesis are based on *in silico* work, which were carried out to understand different properties associated with the mimicry peptides/proteins. When we started this work, there was only one database namely mimicDB which provides information about proteins or epitopes involved in host-pathogen interactions. But it contained predicted information related to only a few human parasites. Henceforth, we foremost developed a database named as miPepBase **(Mi**micry **Pep**tide Data**base),** it incorporated the information related to autoimmune diseases as well as in-depth information about mimicry peptide and proteins. It is freely available at <http://proteininformatics.org/mkumar/mipepbase>.

Though researchers have studied various aspects of molecular mimicry and a few studies have reported some structural characteristics of viral mimicry proteins to the best of our knowledge, structural and functional characteristics of mimicry proteins of viruses, bacteria and hosts have not been explored in detail. Thus, we first time reported the structural and functional characteristics of bacterial, viral and host-mimicry.

The order/disorder propensity in the bacterial, viral and host mimicry proteins/peptides were studied along with the prevalence of intrinsically disordered regions (IDRs) like molecular recognition features (MoRFs), short linear motifs (SLiMs), and low complexity regions (LCRs). Our results indicated that the majority of the bacterial and viral mimicry proteins and mimitopes were ordered and, only a few mimitopes harboured the functional units present in the disordered regions of proteins like MoRFs, SliMs and LCRs. The majority of the host mimicry proteins were disordered, but the host mimitopes were ordered. The fact that most of the host mimitopes were ordered suggests that bacteria and viruses might preferentially select those regions of the host proteins for molecular mimicry which are ordered and thus show a lesser structural flexibility and immunological activity. Functional analyses indicated that both bacterial and viral mimicry proteins were involved in similar functions while the host mimicry proteins were multifunctional and mainly involved in ion binding, symbiont processes and signalling pathways. Our in-depth structure-function relationships in bacteria and viruses can help find ways to mitigate the effects of the infection.

Herein, we first time explained the immunological consequences of bacteria and virus host/pathogen memitopes. The immunogenic study revealed that viral, bacterial, and host memitopes are non-immunogenic, which help them to escape host’s immune response (elaborate..).

Most commonly used therapy for autoimmune diseases has been the usage of immune-suppressants or immune-modulators that treat symptoms rather than the etiology and/or the causative mechanism(s). Although, in many cases (e.g. *Mycobacterium tuberculosis*) these therapies rebound to the host and enhance pathogenesis. In order to overcome this challenge, we developed a novel drug-repurposing methodology to target autoimmune diseases. In which, we ‘tried-and-tested’ our approach to target pathogens at the root cause of disease. Herein, using the interaction partners of pathogen’s mimicry protein as a chokepoint target, we are able identify many potential drug-target molecules for two model organisms namely, *Mycobacterium tuberculosis* complex (MTC)and *Mycobacterium avium* subsp. *paratuberculosis* (MAP). As we found in this work most of the proposed drugs are either FDA-approved or experimentally validated. Henceforth, these can be easily incorporated into clinical studies or tested *in vitro* for assessing their suitability in some autoimmune diseases namely multiple sclerosis, type 1 diabetes mellitus, crohn’s disease and leprosy, caused by MTC and MAP. We also anticipate that the proposed schema can be used to target other pathogens, and number of drug-targets can be increased by incorporating more databases.

In our study, we found that most of the mycobacterium drug target molecules predicted to have a central role in the regulation of latency. These observations encourage us to analyze more in-depth about MTC molecular mimicry phenomenon. Hence, in this thesis using exosomal RNA-seq data of clinical samples collected from active- and latent-TB infected individuals, we explored the pathways in which mimicry proteins and its interacting partners of *Mycobacterium tuberculosis* H37Rv are involved. We observed that majority of mimicry proteins and its interacting partner in the LTBI exosomes that might carry out function far off from the host cell of the exosome. Furthermore, pathway and functional analysis of differentially expressed genes (DEGs) in LTBI and ATB samples indicates a down-regulation of signaling and immune system pathways, and up-regulation of apoptosis/necrosis process. Recent evidence suggests that extracellular vesicles can mediate immune stimulation or suppression and they can drive inflammatory, autoimmune and infectious disease pathology. Thus, modulation of extracellular vesicles has the potential to be used as therapeutic agents to.

# References

[Adam AJ, Schweitzer AN, Sharpe AH. 1998. The role of B7 co-stimulation in activation and differentiation ofCD4+and CD8+T cells. *Immunol. Rev.* 165(1):231–47](http://paperpile.com/b/a3wH5Q/X4JIu)

[Ahmed AR, Spigelman Z, Cavacini LA, Posner MR. 2006. Treatment of pemphigus vulgaris with rituximab and intravenous immune globulin. *N. Engl. J. Med.* 355(17):1772–79](http://paperpile.com/b/a3wH5Q/miCpg)

[Ahmed R. 1992. Immunological memory against viruses. *Semin. Immunol.* 4(2):105–9](http://paperpile.com/b/a3wH5Q/x9X8D)

[Aimanianda V, Haensler J, Lacroix-Desmazes S, Kaveri SV, Bayry J. 2009. Novel cellular and molecular mechanisms of induction of immune responses by aluminum adjuvants. *Trends Pharmacol. Sci.* 30(6):287–95](http://paperpile.com/b/a3wH5Q/c91Ou)

[Albert LJ, Inman RD. 1999. Molecular mimicry and autoimmunity. *N. Engl. J. Med.* 341(27):2068–74](http://paperpile.com/b/a3wH5Q/xt5CD)

[Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402](http://paperpile.com/b/a3wH5Q/97kJM)

[Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, et al. 2005. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* 272(20):5101–9](http://paperpile.com/b/a3wH5Q/QiN6Y)

[Amineni U, Pradhan D, Marisetty H. 2010. In silico identification of common putative drug targets in Leptospira interrogans. *J. Chem. Biol.* 3(4):165–73](http://paperpile.com/b/a3wH5Q/1nKXz)

[Anand P, Puranik A, Aravamudan M, Venkatakrishnan AJ, Soundararajan V. SARS-CoV-2 selectively mimics a cleavable peptide of human ENaC in a strategic hijack of host proteolytic machinery](http://paperpile.com/b/a3wH5Q/OTIvI)

[Ang CW, Jacobs BC, Laman JD. 2004. The Guillain-Barré syndrome: a true case of molecular mimicry. *Trends Immunol.* 25(2):61–66](http://paperpile.com/b/a3wH5Q/MfCtP)

[Anishetty S, Pulimi M, Pennathur G. 2005. Potential drug targets in Mycobacterium tuberculosis through metabolic pathway analysis](http://paperpile.com/b/a3wH5Q/oeQAL)

[An SY, Ka D, Kim I, Kim E-H, Kim N-K, et al. 2020. Intrinsic disorder is essential for Cas9 inhibition of anti-CRISPR AcrIIA5. *Nucleic Acids Res.* 48(13):7584–94](http://paperpile.com/b/a3wH5Q/PEpUZ)

[Appelmelk BJ, Simoons-Smit I, Negrini R, Moran AP, Aspinall GO, et al. 1996. Potential role of molecular mimicry between Helicobacter pylori lipopolysaccharide and host Lewis blood group antigens in autoimmunity. *Infect. Immun.* 64(6):2031–40](http://paperpile.com/b/a3wH5Q/MECL4)

[Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32(Database issue):D115–19](http://paperpile.com/b/a3wH5Q/3tfSK)

[Arru G, Caggiu E, Paulus K, Sechi GP, Mameli G, Sechi LA. 2016. Is there a role for Mycobacterium avium subspecies paratuberculosis in Parkinson’s disease? *J. Neuroimmunol.* 293:86–90](http://paperpile.com/b/a3wH5Q/rwmbu)

[Ascherio A, Munger KL. 2007. Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors. *Ann. Neurol.* 61(6):504–13](http://paperpile.com/b/a3wH5Q/IoucR)

[Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25(1):25–29](http://paperpile.com/b/a3wH5Q/hQYgl)

[Balandraud N, Roudier J, Roudier C. 2004. Epstein-Barr virus and rheumatoid arthritis. *Autoimmun. Rev.* 3(5):362–67](http://paperpile.com/b/a3wH5Q/wIK1H)

[Baltimore D. 1971. Expression of animal virus genomes](http://paperpile.com/b/a3wH5Q/W5j2o)

[Banki K, Colombo E, Sia F, Halladay D, Mattson DH, et al. 1994. Oligodendrocyte-specific expression and autoantigenicity of transaldolase in multiple sclerosis. *J. Exp. Med.* 180(5):1649–63](http://paperpile.com/b/a3wH5Q/e9Bts)

[Barnett LA, Fujinami RS. 1992. Molecular mimicry: a mechanism for autoimmune injury 1](http://paperpile.com/b/a3wH5Q/xUgMP)

[Benkert P, Biasini M, Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 27(3):343–50](http://paperpile.com/b/a3wH5Q/F0MzQ)

[Benkert P, Tosatto SCE, Schomburg D. 2008. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*. 71(1):261–77](http://paperpile.com/b/a3wH5Q/S54eo)

[Benvenga S, Guarneri F. 2016. Molecular mimicry and autoimmune thyroid disease. *Rev. Endocr. Metab. Disord.* 17(4):485–98](http://paperpile.com/b/a3wH5Q/x0vPM)

[Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28(1):235–42](http://paperpile.com/b/a3wH5Q/d2fYk)

[Bertsias GK, Salmon JE, Boumpas DT. 2010. Therapeutic opportunities in systemic lupus erythematosus: state of the art and prospects for the new decade. *Ann. Rheum. Dis.* 69(9):1603–11](http://paperpile.com/b/a3wH5Q/BcEsg)

[Bettelli E, Korn T, Oukka M, Kuchroo VK. 2008. Induction and effector functions of TH17 cells](http://paperpile.com/b/a3wH5Q/eJxrB)

[Biank V, Broeckel U, Kugathasan S. 2007. Pediatric inflammatory bowel disease: Clinical and molecular genetics](http://paperpile.com/b/a3wH5Q/Uy7Qp)

[Binns D, Dimmer E, Huntley R, Barrell D, O’Donovan C, Apweiler R. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*. 25(22):3045–46](http://paperpile.com/b/a3wH5Q/RuPb6)

[Bitti MLM, Masala S, Capasso F, Rapini N, Piccinini S, et al. 2012. Mycobacterium avium subsp. paratuberculosis in an Italian cohort of type 1 diabetes pediatric patients. *Clin. Dev. Immunol.* 2012:785262](http://paperpile.com/b/a3wH5Q/YdXqh)

[Bo M, Erre GL, Niegowska M, Piras M, Taras L, et al. 2018. Interferon regulatory factor 5 is a potential target of autoimmune response triggered by Epstein-barr virus and Mycobacterium avium subsp. paratuberculosis in rheumatoid arthritis: investigating a mechanism of molecular mimicry. *Clin. Exp. Rheumatol.* 36(3):376–81](http://paperpile.com/b/a3wH5Q/5z9aL)

[Bouvrette LPB, Cody NAL, Bergalet J, Lefebvre FA, Diot C, et al. 2018. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells](http://paperpile.com/b/a3wH5Q/PkFTy)

[Bridges SL. 2004. Update on autoantibodies in rheumatoid arthritis. *Curr. Rheumatol. Rep.* 6(5):343–50](http://paperpile.com/b/a3wH5Q/CU8sp)

[Burg JS, Ingram JR, Venkatakrishnan AJ, Jude KM, Dukkipati A, et al. 2015. Structural biology. Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor. *Science*. 347(6226):1113–17](http://paperpile.com/b/a3wH5Q/JnGM)

[Ceccaldi P, Rendon J, Léger C, Toci R, Guigliarelli B, et al. 2015. Reductive activation of E. coli respiratory nitrate reductase. *Biochim. Biophys. Acta*. 1847(10):1055–63](http://paperpile.com/b/a3wH5Q/GgePA)

[Chamberlin W, Borody TJ, Campbell J. 2011. Primary treatment of Crohn’s disease: combined antibiotics taking center stage. *Expert Rev. Clin. Immunol.* 7(6):751–60](http://paperpile.com/b/a3wH5Q/DkDab)

[Chandrashekara S. 2012. The treatment strategies of autoimmune disease may need a different approach from conventional protocol: A review](http://paperpile.com/b/a3wH5Q/gJrjO)

[Chan SL, Low LY, Hsu S, Li S, Liu T, et al. 2009. Molecular mimicry in innate immunity: crystal structure of a bacterial TIR domain. *J. Biol. Chem.* 284(32):21386–92](http://paperpile.com/b/a3wH5Q/athA)

[Chaplin DD. 2003. 1. Overview of the immune response. *J. Allergy Clin. Immunol.* 111(2 Suppl):S442–59](http://paperpile.com/b/a3wH5Q/M39hT)

[Chen CR, Tanaka K, Chazenbalk GD, McLachlan SM, Rapoport B. 2001. A full biological response to autoantibodies in Graves’ disease requires a disulfide-bonded loop in the thyrotropin receptor N terminus homologous to a laminin epidermal growth factor-like domain. *J. Biol. Chem.* 276(18):14767–72](http://paperpile.com/b/a3wH5Q/h9jhK)

[Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, et al. 2010. MolProbity : all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66(1):12–21](http://paperpile.com/b/a3wH5Q/WDbW4)

[Chen W, Feng P, Ding H, Lin H, Chou K-C. 2015. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* 490:26–33](http://paperpile.com/b/a3wH5Q/0kLtC)

[Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C. 2014. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* 456:53–60](http://paperpile.com/b/a3wH5Q/mtR5f)

[Chong C-E, Lim B-S, Nathan S, Mohamed R. 2006. In silico analysis of Burkholderia pseudomallei genome sequence for potential drug targets. *In Silico Biol.* 6(4):341–46](http://paperpile.com/b/a3wH5Q/UaM2C)

[Christen U. 2019. Pathogen infection and autoimmune disease. *Clin. Exp. Immunol.* 195(1):10–14](http://paperpile.com/b/a3wH5Q/lFQlc)

[Christen U, Edelmann KH, McGavern DB, Wolfe T, Coon B, et al. 2004. A viral epitope that mimics a self antigen can accelerate but not initiate autoimmune diabetes. *J. Clin. Invest.* 114(9):1290–98](http://paperpile.com/b/a3wH5Q/5UUv5)

[Christen U, Hintermann E. 2018. Autoantibodies in Autoimmune Hepatitis: Can Epitopes Tell Us about the Etiology of the Disease? *Front. Immunol.* 9:163](http://paperpile.com/b/a3wH5Q/x1vi7)

[Chung BK-S, Dick T, Lee D-Y. 2013. In silico analyses for the discovery of tuberculosis drug targets. *J. Antimicrob. Chemother.* 68(12):2701–9](http://paperpile.com/b/a3wH5Q/sjg4H)

[Cirl C, Wieser A, Yadav M, Duerr S, Schubert S, et al. 2008. Subversion of Toll-like receptor signaling by a unique family of bacterial Toll/interleukin-1 receptor domain-containing proteins. *Nat. Med.* 14(4):399–406](http://paperpile.com/b/a3wH5Q/hY93)

[Clayton K, Polak ME, Woelk CH, Elkington P. 2017. Gene Expression Signatures in Tuberculosis Have Greater Overlap with Autoimmune Diseases Than with Infectious Diseases. *Am. J. Respir. Crit. Care Med.* 196(5):655–56](http://paperpile.com/b/a3wH5Q/EOvqr)

[Coletta A, Pinney JW, Solís DYW, Marsh J, Pettifer SR, Attwood TK. 2010. Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst. Biol.* 4(1):19](http://paperpile.com/b/a3wH5Q/nygH3)

[Cooper GS, Miller FW, Pandey JP. 1999. The role of genetic factors in autoimmune disease: implications for environmental research. *Environ. Health Perspect.* 107(suppl 5):693–700](http://paperpile.com/b/a3wH5Q/e9Ve1)

[Coppieters KT, Wiberg A, von Herrath MG. 2012. Viral infections and molecular mimicry in type 1 diabetes. *APMIS*. 120(12):941–49](http://paperpile.com/b/a3wH5Q/AW4dN)

[Cossu D, Mameli G, Masala S, Cocco E, Frau J, et al. 2014. Evaluation of the humoral response against mycobacterial peptides, homologous to MOG35–55, in multiple sclerosis patients](http://paperpile.com/b/a3wH5Q/XC2FW)

[Cossu D, Masala S, Cocco E, Paccagnini D, Tranquilli S, et al. 2013. Association of Mycobacterium avium subsp. paratuberculosis and SLC11A1 polymorphisms in Sardinian multiple sclerosis patients. *J. Infect. Dev. Ctries.* 7(3):203–7](http://paperpile.com/b/a3wH5Q/OodZQ)

[Cossu D, Yokoyama K, Hattori N. 2017. Conflicting Role of Species in Multiple Sclerosis. *Front. Neurol.* 8:216](http://paperpile.com/b/a3wH5Q/38sGN)

[Cossu D, Yokoyama K, Hattori N. 2018. Bacteria-Host Interactions in Multiple Sclerosis. *Front. Microbiol.* 9:2966](http://paperpile.com/b/a3wH5Q/B74kP)

[Cunha-Neto E, Kalil J. 2014. Molecular Mimicry and Chagas’ Disease](http://paperpile.com/b/a3wH5Q/j9yBt)

[Cusick MF, Libbey JE, Fujinami RS. 2012. Molecular mimicry as a mechanism of autoimmune disease. *Clin. Rev. Allergy Immunol.* 42(1):102–11](http://paperpile.com/b/a3wH5Q/WvBKB)

[Damian RT. 1964. Molecular Mimicry: Antigen Sharing by Parasite and Host and Its Consequences](http://paperpile.com/b/a3wH5Q/gmerF)

[D’Amore M, Lisi S, Sisto M, Cucci L, Thomas Dow C. 2010. Molecular identification of Mycobacterium avium subspecies paratuberculosis in an Italian patient with Hashimoto’s thyroiditis and Melkersson–Rosenthal syndrome](http://paperpile.com/b/a3wH5Q/VFr3x)

[Davey NE, Travé G, Gibson TJ. 2011. How viruses hijack cell regulation. *Trends Biochem. Sci.* 36(3):159–69](http://paperpile.com/b/a3wH5Q/EpT9z)

[Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, et al. 2012. Attributes of short linear motifs. *Mol. Biosyst.* 8(1):268–81](http://paperpile.com/b/a3wH5Q/d0tzO)

[Davies JM. 1997. Molecular mimicry: Can epitope mimicry induce autoimmune disease?](http://paperpile.com/b/a3wH5Q/qwIIp)

[Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, et al. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* 35(Web Server issue):W375–83](http://paperpile.com/b/a3wH5Q/IdHfN)

[Deiana A, Forcelloni S, Porrello A, Giansanti A. 2018. *New classification of intrinsic disorder in the Human proteome*. Work. Pap.](http://paperpile.com/b/a3wH5Q/nQ3z7)

[D’Elios MM, Appelmelk BJ, Amedei A, Bergman MP, Del Prete G. 2004. Gastric autoimmunity: the role of Helicobacter pylori and molecular mimicry](http://paperpile.com/b/a3wH5Q/slqTZ)

[Diaz PI, Hong B-Y, Frias-Lopez J, Dupuy AK, Angeloni M, et al. 2013. Transplantation-associated long-term immunosuppression promotes oral colonization by potentially opportunistic pathogens without impacting other members of the salivary bacteriome. *Clin. Vaccine Immunol.* 20(6):920–30](http://paperpile.com/b/a3wH5Q/LsSVj)

[Dinarello CA. 2003. Anti-cytokine therapeutics and infections](http://paperpile.com/b/a3wH5Q/4AKBT)

[Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, et al. 2014. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* 42(Database issue):D259–66](http://paperpile.com/b/a3wH5Q/38qz6)

[Dolan PT, Roth AP, Xue B, Sun R, Dunker AK, et al. 2015. Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions: HCV Core MoRFs Target Cellular Proteins. *Protein Sci.* 24(2):221–35](http://paperpile.com/b/a3wH5Q/kVaOV)

[Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 21(16):3433–34](http://paperpile.com/b/a3wH5Q/zPp8q)

[Dosztányi Z, Mészáros B, Simon I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*. 25(20):2745–46](http://paperpile.com/b/a3wH5Q/M9Mm1)

[Dow CT. 2006. Paratuberculosis and Type I diabetes: is this the trigger? *Med. Hypotheses*. 67(4):782–85](http://paperpile.com/b/a3wH5Q/wHNaH)

[Dow CT. 2012. M. paratuberculosis Heat Shock Protein 65 and Human Diseases: Bridging Infection and Autoimmunity. *Autoimmune Dis.* 2012:150824](http://paperpile.com/b/a3wH5Q/dfBKf)

[Dow CT. 2014. M. paratuberculosis and Parkinson’s disease – Is this a trigger](http://paperpile.com/b/a3wH5Q/gFV1F)

[Dow CT, Ellingson JLE. 2010. Detection of Mycobacterium avium ss. Paratuberculosis in Blau Syndrome Tissues. *Autoimmune Dis.* 2011:127692](http://paperpile.com/b/a3wH5Q/3hyI5)

[Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. 2010. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics*. 11:222](http://paperpile.com/b/a3wH5Q/Aka8R)

[Drayman N, Glick Y, Ben-nun-Shaul O, Zer H, Zlotnick A, et al. 2013. Pathogens Use Structural Mimicry of Native Host Ligands as a Mechanism for Host Receptor Engagement. *Cell Host Microbe*. 14(1):63–73](http://paperpile.com/b/a3wH5Q/mpgRS)

[Duffield M, Cooper I, McAlister E, Bayliss M, Ford D, Oyston P. 2010. Predicting conserved essential genes in bacteria: in silico identification of putative drug targets. *Mol. Biosyst.* 6(12):2482–89](http://paperpile.com/b/a3wH5Q/A9jFn)

[Duncan K. 2004. Identification and validation of novel drug targets in tuberculosis. *Curr. Pharm. Des.* 10(26):3185–94](http://paperpile.com/b/a3wH5Q/OfhCO)

[Duro N, Miskei M, Fuxreiter M. 2015. Fuzziness endows viral motif-mimicry. *Mol. Biosyst.* 11(10):2821–29](http://paperpile.com/b/a3wH5Q/vZh5g)

[Du Z, Uversky VN. 2017. Functional roles of intrinsic disorder in CRISPR-associated protein Cas9](http://paperpile.com/b/a3wH5Q/f1QDX)

[Ebringer A, Rashid T. 2009. Rheumatoid arthritis is caused by Proteus: the molecular mimicry theory and Karl Popper. *Front. Biosci.* . 1:577–86](http://paperpile.com/b/a3wH5Q/5MZiN)

[Elkington P, Tebruegge M, Mansour S. 2016. Tuberculosis: An Infection-Initiated Autoimmune Disease? *Trends Immunol.* 37(12):815–18](http://paperpile.com/b/a3wH5Q/HZU7I)

[Farris AD, Keech CL, Gordon TP, McCluskey J. 2000. Epitope mimics and determinant spreading: pathways to autoimmunity. *Cell. Mol. Life Sci.* 57(4):569–78](http://paperpile.com/b/a3wH5Q/MlDsU)

[Finlay BB, McFadden G. 2006. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*. 124(4):767–82](http://paperpile.com/b/a3wH5Q/DMXqx)

[Franzosa EA, Garamszegi S, Xia Y. 2012. Toward a three-dimensional view of protein networks between species. *Front. Microbiol.* 3:428](http://paperpile.com/b/a3wH5Q/VjOj)

[Franzosa EA, Xia Y. 2011. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U. S. A.* 108(26):10538–43](http://paperpile.com/b/a3wH5Q/gjiQ)

[Fridkis-Hareli M. 2008. Immunogenetic mechanisms for the coexistence of organ-specific and systemic autoimmune diseases. *J. Autoimmune Dis.* 5:1](http://paperpile.com/b/a3wH5Q/TmBcm)

[Fujinami RS. 2001. Viruses and autoimmune disease – two sides of the same coin? *Trends Microbiol.* 9(8):377–81](http://paperpile.com/b/a3wH5Q/olgUf)

[Fujinami RS, Oldstone MB. 1985. Amino acid homology between the encephalitogenic site of myelin basic protein and virus: mechanism for autoimmunity. *Science*. 230(4729):1043–45](http://paperpile.com/b/a3wH5Q/jgXPb)

[Fujinami RS, Oldstone MB, Wroblewska Z, Frankel ME, Koprowski H. 1983. Molecular mimicry in virus infection: crossreaction of measles virus phosphoprotein or of herpes simplex virus protein with human intermediate filaments. *Proc. Natl. Acad. Sci. U. S. A.* 80(8):2346–50](http://paperpile.com/b/a3wH5Q/LBBq9)

[Fujinami RS, von Herrath MG, Christen U, Lindsay Whitton J. 2006. Molecular Mimicry, Bystander Activation, or Viral Persistence: Infections and Autoimmune Disease](http://paperpile.com/b/a3wH5Q/EQ17z)

[Garg A, Kumari B, Kumar R, Kumar M. 2017. miPepBase: A Database of Experimentally Verified Peptides Involved in Molecular Mimicry. *Front. Microbiol.* 8:2053](http://paperpile.com/b/a3wH5Q/ElOF5)

[Garg A, Kumari B, Singhal N, Kumar M. 2019. Using molecular-mimicry-inducing pathways of pathogens as novel drug targets. *Drug Discov. Today*. 24(9):1943–52](http://paperpile.com/b/a3wH5Q/CJqgW)

[Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(Database issue):D1100–1107](http://paperpile.com/b/a3wH5Q/eenA1)

[Gauntt CJ, Tracy SM, Chapman N, Wood HJ, Kolbeck PC, et al. 1995. Coxsackievirus-induced chronic myocarditis in murine models. *Eur. Heart J.* 16 Suppl O:56–58](http://paperpile.com/b/a3wH5Q/1J254)

[Gomez JE, McKinney JD. 2004. M. tuberculosis persistence, latency, and drug tolerance. *Tuberculosis* . 84(1-2):29–44](http://paperpile.com/b/a3wH5Q/mHyPl)

[Gong L, Liu B, Wang J, Pan H, Qi A, et al. 2018. Novel missense mutation in PTPN22 in a Chinese pedigree with Hashimoto’s thyroiditis](http://paperpile.com/b/a3wH5Q/4kIa2)

[Gowthaman U, Eswarakumar VP. 2013. Molecular mimicry: good artists copy, great artists steal. *Virulence*. 4(6):433–34](http://paperpile.com/b/a3wH5Q/p4TgP)

[Greenstein RJ, Su L, Brown ST. 2010. The thioamides methimazole and thiourea inhibit growth of M. avium Subspecies paratuberculosis in culture. *PLoS One*. 5(6):e11099](http://paperpile.com/b/a3wH5Q/KmcP7)

[Greenstein RJ, Su L, Juste RA, Brown ST. 2008. On the Action of Cyclosporine A, Rapamycin and Tacrolimus on M. avium Including Subspecies paratuberculosis](http://paperpile.com/b/a3wH5Q/j1CSy)

[Gregersen PK, Silver J, Winchester RJ. 1987. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* 30(11):1205–13](http://paperpile.com/b/a3wH5Q/doQiB)

[Grossman C, Dovrish Z, Shoenfeld Y, Amital H. 2011. Do infections facilitate the emergence of systemic sclerosis? *Autoimmun. Rev.* 10(5):244–47](http://paperpile.com/b/a3wH5Q/SXpMn)

[Guarneri F. 2013. Molecular mimicry in cutaneous autoimmune diseases. *WJD*. 2(4):36](http://paperpile.com/b/a3wH5Q/OwHVO)

[Gupta M, Prasad Y, Sharma SK, Jain CK. 2017. Identification of Phosphoribosyl-AMP cyclohydrolase, as drug target and its inhibitors in Brucella melitensis bv. 1 16M using metabolic pathway analysis. *J. Biomol. Struct. Dyn.* 35(2):287–99](http://paperpile.com/b/a3wH5Q/qIRqg)

[Guven Maiorov E, Keskin O, Gursoy A, Nussinov R. 2013. The structural network of inflammation and cancer: merits and challenges. *Semin. Cancer Biol.* 23(4):243–51](http://paperpile.com/b/a3wH5Q/tMPO)

[Guven-Maiorov E, Keskin O, Gursoy A, VanWaes C, Chen Z, et al. 2015. The Architecture of the TIR Domain Signalosome in the Toll-like Receptor-4 Signaling Pathway. *Sci. Rep.* 5:13128](http://paperpile.com/b/a3wH5Q/dHDA)

[Guven-Maiorov E, Tsai C-J, Nussinov R. 2016. Pathogen mimicry of host protein-protein interfaces modulates immunity. *Semin. Cell Dev. Biol.* 58:136–45](http://paperpile.com/b/a3wH5Q/gffNr)

[Hagai T, Azia A, Babu MM, Andino R. 2014. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Rep.* 7(5):1729–39](http://paperpile.com/b/a3wH5Q/JLIZ)

[Halehalli RR, Nagarajaram HA. 2015. Molecular principles of human virus protein-protein interactions. *Bioinformatics*. 31(7):1025–33](http://paperpile.com/b/a3wH5Q/kFS5f)

[Harbige J, Eichmann M, Peakman M. 2017. New insights into non-conventional epitopes as T cell targets: The missing link for breaking immune tolerance in autoimmune disease? *J. Autoimmun.* 84:12–20](http://paperpile.com/b/a3wH5Q/FhP55)

[Hardtke-Wolenski M, Dywicki J, Fischer K, Hapke M, Sievers M, et al. 2017. The influence of genetic predisposition and autoimmune hepatitis inducing antigens in disease development. *J. Autoimmun.* 78:39–45](http://paperpile.com/b/a3wH5Q/Fom9p)

[Härtlova A, Herbst S, Peltier J, Rodgers A, Bilkei‐Gorzo O, et al. 2018. LRRK2 is a negative regulator of Mycobacterium tuberculosis phagosome maturation in macrophages](http://paperpile.com/b/a3wH5Q/f87Pq)

[Heasman J, Wessely O, Langland R, Craig EJ, Kessler DS. 2001. Vegetal localization of maternal mRNAs is disrupted by VegT depletion. *Dev. Biol.* 240(2):377–86](http://paperpile.com/b/a3wH5Q/wxiLp)

[Hentschel J, Burnside C, Mignot I, Leibundgut M, Boehringer D, Ban N. 2017. The Complete Structure of the Mycobacterium smegmatis 70S Ribosome. *Cell Rep.* 20(1):149–60](http://paperpile.com/b/a3wH5Q/PcML0)

[Hiemstra HS, Schloot NC, van Veelen PA, Willemen SJ, Franken KL, et al. 2001. Cytomegalovirus in autoimmunity: T cell crossreactivity to viral antigen and autoantigen glutamic acid decarboxylase. *Proc. Natl. Acad. Sci. U. S. A.* 98(7):3988–91](http://paperpile.com/b/a3wH5Q/vkFzp)

[Ho BK, Brasseur R. 2005. The Ramachandran plots of glycine and pre-proline. *BMC Struct. Biol.* 5:14](http://paperpile.com/b/a3wH5Q/tkTEv)

[Hraber P, O’Maille PE, Silberfarb A, Davis-Anderson K, Generous N, et al. 2020. Resources to Discover and Use Short Linear Motifs in Viral Proteins. *Trends Biotechnol.* 38(1):113–27](http://paperpile.com/b/a3wH5Q/dy5XJ)

[Huber SA, Moraska A, Cunningham M. 1994. Alterations in major histocompatibility complex association of myocarditis induced by coxsackievirus B3 mutants selected with monoclonal antibodies to group A streptococci. *Proc. Natl. Acad. Sci. U. S. A.* 91(12):5543–47](http://paperpile.com/b/a3wH5Q/H9sKV)

[Hughes SC, Simmonds AJ. 2019. Drosophila mRNA Localization During Later Development: Past, Present, and Future](http://paperpile.com/b/a3wH5Q/0zS9I)

[Hung AW, Silvestre HL, Wen S, George GPC, Boland J, et al. 2016. Optimization of Inhibitors of Mycobacterium tuberculosis Pantothenate Synthetase Based on Group Efficiency Analysis. *ChemMedChem*. 11(1):38–42](http://paperpile.com/b/a3wH5Q/uqirb)

[Igoe A, Scofield RH. 2013. Autoimmunity and infection in Sjögrenʼs syndrome. *Curr. Opin. Rheumatol.* 25(4):480–87](http://paperpile.com/b/a3wH5Q/CZgTm)

[Ioerger TR, O’Malley T, Liao R, Guinn KM, Hickey MJ, et al. 2013. Identification of New Drug Targets and Resistance Mechanisms in Mycobacterium tuberculosis](http://paperpile.com/b/a3wH5Q/vfXtg)

[Irwin JJ, Shoichet BK. 2005. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45(1):177–82](http://paperpile.com/b/a3wH5Q/gqQ8m)

[Janeway C Jr, Medzhitov R. 2000. Viral interference with IL-1 and toll signaling](http://paperpile.com/b/a3wH5Q/gOVW)

[John Wiley & Sons, Ltd, ed. 2001. Molecular Mimicry. In *Encyclopedia of Life Sciences*, Vol. 15, p. 547. Chichester, UK: John Wiley & Sons, Ltd](http://paperpile.com/b/a3wH5Q/s4tKM)

[Johri AK, Paoletti LC, Glaser P, Dua M, Sharma PK, et al. 2006. Group B Streptococcus: global incidence and vaccine development. *Nat. Rev. Microbiol.* 4(12):932–42](http://paperpile.com/b/a3wH5Q/mlkaN)

[Jonczyk R, Ronconi S, Rychlik M, Genschel U. 2008. Pantothenate synthetase is essential but not limiting for pantothenate biosynthesis in Arabidopsis](http://paperpile.com/b/a3wH5Q/jDzc6)

[Jones DT, Cozzetto D. 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 31(6):857–63](http://paperpile.com/b/a3wH5Q/APHRY)

[Jörg S, Grohme DA, Erzler M, Binsfeld M, Haghikia A, et al. 2016. Environmental factors in autoimmune diseases and their role in multiple sclerosis. *Cell. Mol. Life Sci.* 73(24):4611–22](http://paperpile.com/b/a3wH5Q/Po01p)

[Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, Ting AY. 2017. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife*. 6:](http://paperpile.com/b/a3wH5Q/98Wcb)

[Kakumanu P, Yamagata H, Sobel ES, Reeves WH, Chan EKL, Satoh M. 2008. Patients with pulmonary tuberculosis are frequently positive for anti-cyclic citrullinated peptide antibodies, but their sera also react with unmodified arginine-containing peptide. *Arthritis Rheum.* 58(6):1576–81](http://paperpile.com/b/a3wH5Q/KrpDm)

[Kanehisa M. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes](http://paperpile.com/b/a3wH5Q/nrKEf)

[Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45(D1):D353–61](http://paperpile.com/b/a3wH5Q/jChxW)

[Kaplan MH, Meyeserian M. 1962. An immunological cross-reaction between group-A streptococcal cells and human heart tissue. *Lancet*. 1(7232):706–10](http://paperpile.com/b/a3wH5Q/rDKkO)

[Kaufman KM, Kirby MY, Harley JB, James JA. 2003. Peptide mimics of a major lupus epitope of SmB/B’. *Ann. N. Y. Acad. Sci.* 987:215–29](http://paperpile.com/b/a3wH5Q/4LeCk)

[Kazatchkine MD, Kaveri SV. 2001. Immunomodulation of autoimmune and inflammatory diseases with intravenous immune globulin. *N. Engl. J. Med.* 345(10):747–55](http://paperpile.com/b/a3wH5Q/DoLtG)

[Keskin O, Nussinov R. 2005. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng. Des. Sel.* 18(1):11–24](http://paperpile.com/b/a3wH5Q/AZT1)

[Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, et al. 2016. PubChem Substance and Compound databases. *Nucleic Acids Res.* 44(D1):D1202–13](http://paperpile.com/b/a3wH5Q/XVirO)

[Kivity S, Agmon-Levin N, Blank M, Shoenfeld Y. 2009. Infections and autoimmunity--friends or foes? *Trends Immunol.* 30(8):409–14](http://paperpile.com/b/a3wH5Q/DQDnT)

[Kloc M, Etkin LD. 1994. Delocalization of Vg1 mRNA from the vegetal cortex in Xenopus oocytes after destruction of Xlsirt RNA. *Science*. 265(5175):1101–3](http://paperpile.com/b/a3wH5Q/ranaK)

[Kloc M, Ruth Zearfoss N, Etkin LD. 2002. Mechanisms of Subcellular mRNA Localization](http://paperpile.com/b/a3wH5Q/7xmtf)

[Kohn LD, Napolitano G, Singer DS, Molteni M, Scorza R, et al. 2000. Graves’ disease: a host defense mechanism gone awry. *Int. Rev. Immunol.* 19(6):633–64](http://paperpile.com/b/a3wH5Q/yDzXo)

[Kolls JK, Khader SA. 2011. TH17 Cytokines in Primary Mucosal Immunity](http://paperpile.com/b/a3wH5Q/LXY7D)

[Konforte D, Diamandis EP, van Venrooij WJ, Lories R, Ward MM. 2012. Autoimmune diseases: early diagnosis and new treatment strategies. *Clin. Chem.* 58(11):1510–14](http://paperpile.com/b/a3wH5Q/0FKzY)

[Kumar M, Raghava GPS. 2009. Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics*. 10:22](http://paperpile.com/b/a3wH5Q/y2ntf)

[Kumar M, Verma R, Raghava GPS. 2006. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J. Biol. Chem.* 281(9):5357–63](http://paperpile.com/b/a3wH5Q/4SAH9)

[Kumar R, Jain S, Kumari B, Kumar M. 2014. Protein sub-nuclear localization prediction using SVM and Pfam domain information. *PLoS One*. 9(6):e98345](http://paperpile.com/b/a3wH5Q/KSNgL)

[Kumar R, Kumari B, Kumar M. 2018. Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. *Mitochondrion*. 42:11–22](http://paperpile.com/b/a3wH5Q/Naxum)

[Lamichhane G. 2011. Novel targets in M. tuberculosis: search for new drugs](http://paperpile.com/b/a3wH5Q/lMYtJ)

[Laxminarayana D. 2017. Is Tolerance Broken in Autoimmunity? *Clin. Med. Insights Pathol.* 10:1179555717712716](http://paperpile.com/b/a3wH5Q/sgTse)

[Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*. 131(1):174–87](http://paperpile.com/b/a3wH5Q/Tp6h6)

[Lester PJ, Buick KH, Baty JW, Felden A, Haywood J. 2019. Different bacterial and viral pathogens trigger distinct immune responses in a globally invasive ant. *Sci. Rep.* 9(1):5780](http://paperpile.com/b/a3wH5Q/UD9r6)

[Libbey JE, Fujinami RS. 2010. Potential triggers of MS. *Results Probl. Cell Differ.* 51:21–42](http://paperpile.com/b/a3wH5Q/vdVQ2)

[Libbey JE, McCoy LL, Fujinami RS. 2007. Molecular Mimicry in Multiple Sclerosis. In *The Neurobiology of Multiple Sclerosis*, Vol. 79, pp. 127–47. Elsevier](http://paperpile.com/b/a3wH5Q/nODs2)

[Liegro CMDI, Di Liegro CM, Schiera G, Di Liegro I. 2014. Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review)](http://paperpile.com/b/a3wH5Q/aVIcI)

[Liu B, Fang L, Long R, Lan X, Chou K-C. 2016. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*. 32(3):362–69](http://paperpile.com/b/a3wH5Q/gzmWz)

[Liu B, Liu F, Fang L, Wang X, Chou K-C. 2015. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*. 31(8):1307–9](http://paperpile.com/b/a3wH5Q/sTjIY)

[Liu B, Wang S, Long R, Chou K-C. 2017a. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*. 33(1):35–41](http://paperpile.com/b/a3wH5Q/WLwBU)

[Liu B, Yang F, Chou K-C. 2017b. 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. *Mol. Ther. Nucleic Acids*. 7:267–77](http://paperpile.com/b/a3wH5Q/CR7z0)

[Liu B, Yang F, Huang D-S, Chou K-C. 2018. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics*. 34(1):33–40](http://paperpile.com/b/a3wH5Q/vW8bf)

[Liu H, Wei Q, Huang C, Zhang Y, Guo Z. 2017c. Potential Roles of Intrinsic Disorder in Maternal-Effect Proteins Involved in the Maintenance of DNA Methylation. *Int. J. Mol. Sci.* 18(9):](http://paperpile.com/b/a3wH5Q/lHxki)

[Li YH, Yu CY, Li XX, Zhang P, Tang J, et al. 2018. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics](http://paperpile.com/b/a3wH5Q/PNEg3)

[Ludin P, Nilsson D, Mäser P. 2011. Genome-wide identification of molecular mimicry candidates in parasites. *PLoS One*. 6(3):e17546](http://paperpile.com/b/a3wH5Q/OSJm)

[Malhis N, Jacobson M, Gsponer J. 2016. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* 44(W1):W488–93](http://paperpile.com/b/a3wH5Q/Wa6D8)

[Mameli G, Cocco E, Frau J, Marrosu MG, Sechi LA. 2016. Epstein Barr Virus and Mycobacterium avium subsp. paratuberculosis peptides are recognized in sera and cerebrospinal fluid of MS patients. *Sci. Rep.* 6:22401](http://paperpile.com/b/a3wH5Q/XAkIe)

[Mameli G, Cossu D, Cocco E, Masala S, Frau J, et al. 2014. Epstein-Barr virus and Mycobacterium avium subsp. paratuberculosis peptides are cross recognized by anti-myelin basic protein antibodies in multiple sclerosis patients. *J. Neuroimmunol.* 270(1-2):51–55](http://paperpile.com/b/a3wH5Q/uAquW)

[Maródi L, Casanova J-L. 2010. Can primary immunodeficiencies help to provide insights into infectious risks of therapeutic antibodies?](http://paperpile.com/b/a3wH5Q/l5vTJ)

[Mathis D, Vence L, Benoist C. 2001. beta-Cell death during progression to diabetes. *Nature*. 414(6865):792–98](http://paperpile.com/b/a3wH5Q/lvOl7)

[Ma W-T, Chang C, Gershwin ME, Lian Z-X. 2017. Development of autoantibodies precedes clinical manifestations of autoimmune diseases: A comprehensive review. *J. Autoimmun.* 83:95–112](http://paperpile.com/b/a3wH5Q/MN3N4)

[McClain MT, Heinlen LD, Dennis GJ, Roebuck J, Harley JB, James JA. 2005. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. *Nat. Med.* 11(1):85–89](http://paperpile.com/b/a3wH5Q/8vuSq)

[McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32(Web Server issue):W20–25](http://paperpile.com/b/a3wH5Q/TkVk7)

[McNees AL, Markesich D, Zayyani NR, Graham DY. 2015. Mycobacterium paratuberculosis as a cause of Crohn’s disease. *Expert Rev. Gastroenterol. Hepatol.* 9(12):1523–34](http://paperpile.com/b/a3wH5Q/sm84x)

[Mdluli K, Spigelman M. 2006. Novel targets for tuberculosis drug discovery. *Curr. Opin. Pharmacol.* 6(5):459–67](http://paperpile.com/b/a3wH5Q/jufHw)

[Medioni C, Mowry K, Besse F. 2012. Principles and roles of mRNA localization in animal development. *Development*. 139(18):3263–76](http://paperpile.com/b/a3wH5Q/ntQJM)

[Mering C v., v. Mering C. 2003. STRING: a database of predicted functional associations between proteins](http://paperpile.com/b/a3wH5Q/Z65oj)

[Miller FW. 2011. Environmental agents and autoimmune diseases. *Adv. Exp. Med. Biol.* 711:61–81](http://paperpile.com/b/a3wH5Q/br8Ex)

[Mishra NK, Kumar M, Raghava GPS. 2007. Support vector machine based prediction of glutathione S-transferase proteins. *Protein Pept. Lett.* 14(6):575–80](http://paperpile.com/b/a3wH5Q/nr4vS)

[Mishra PM, Verma NC, Rao C, Uversky VN, Nandi CK. 2020. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. *Prog. Mol. Biol. Transl. Sci.* 174:1–78](http://paperpile.com/b/a3wH5Q/5JgU9)

[Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, et al. 2006. Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol.* 362(5):1043–59](http://paperpile.com/b/a3wH5Q/tkb92)

[Montero F, Martínez-Barrio J, Serrano-Benavente B, González T, Rivera J, et al. 2020. Coronavirus disease 2019 (COVID-19) in autoimmune and inflammatory conditions: clinical characteristics of poor outcomes. *Rheumatol. Int.* 40(10):1593–98](http://paperpile.com/b/a3wH5Q/kY6zJ)

[Naser SA, Ghobrial G, Romero C, Valentine JF. 2004. Culture of Mycobacterium avium subspecies paratuberculosis from the blood of patients with Crohn’s disease. *Lancet*. 364(9439):1039–44](http://paperpile.com/b/a3wH5Q/tiJ66)

[Naser SA, Thanigachalam S, Dow CT, Collins MT. 2013. Exploring the role of Mycobacterium avium subspecies paratuberculosis in the pathogenesis of type 1 diabetes mellitus: a pilot study. *Gut Pathog.* 5:14](http://paperpile.com/b/a3wH5Q/r7OT4)

[Nellore A, Fishman JA. 2016. The Microbiome, Systemic Immune Function, and Allotransplantation. *Clin. Microbiol. Rev.* 29(1):191–99](http://paperpile.com/b/a3wH5Q/F9ki1)

[Neumann W, Hadley RC, Nolan EM. 2017. Transition metals at the host-pathogen interface: how exploit human metalloproteins for acquiring iron and zinc. *Essays Biochem.* 61(2):211–23](http://paperpile.com/b/a3wH5Q/30XSo)

[Neu N, Rose NR, Beisel KW, Herskowitz A, Gurri-Glass G, Craig SW. 1987. Cardiac myosin induces myocarditis in genetically predisposed mice. *J. Immunol.* 139(11):3630–36](http://paperpile.com/b/a3wH5Q/QgsvU)

[Ng VH, Cox JS, Sousa AO, MacMicking JD, McKinney JD. 2004. Role of KatG catalase-peroxidase in mycobacterial pathogenesis: countering the phagocyte oxidative burst. *Mol. Microbiol.* 52(5):1291–1302](http://paperpile.com/b/a3wH5Q/iqSsE)

[Nielsen TR, Pedersen M, Rostgaard K, Frisch M, Hjalgrim H. 2007. Correlations between Epstein-Barr virus antibody levels and risk factors for multiple sclerosis in healthy individuals. *Mult. Scler.* 13(3):420–23](http://paperpile.com/b/a3wH5Q/lJ5nD)

[Nisini R. 2016. Targeting Immune-Evasion Mechanisms as a Possible New Approach in the Fight against Tuberculosis](http://paperpile.com/b/a3wH5Q/vzgjk)

[Nussinov R, Ma B, Tsai C-J. 2014. Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys. Chem.* 186:22–30](http://paperpile.com/b/a3wH5Q/R79ms)

[Oldstone MBA. 1998. Molecular mimicry and immune‐mediated diseases](http://paperpile.com/b/a3wH5Q/Hfw6I)

[Oldstone MBA. 2005. Molecular mimicry, microbial infection, and autoimmune disease: evolution of the concept. *Curr. Top. Microbiol. Immunol.* 296:1–17](http://paperpile.com/b/a3wH5Q/PEA8y)

[O’Shea C, Staby L, Bendsen SK, Tidemand FG, Redsted A, et al. 2017. Structures and Short Linear Motif of Disordered Transcription Factor Regions Provide Clues to the Interactome of the Cellular Hub Protein Radical-induced Cell Death1. *J. Biol. Chem.* 292(2):512–27](http://paperpile.com/b/a3wH5Q/MaKKD)

[Paccagnini D, Sieswerda L, Rosu V, Masala S, Pacifico A, et al. 2009. Linking chronic infection and autoimmune diseases: Mycobacterium avium subspecies paratuberculosis, SLC11A1 polymorphisms and type-1 diabetes mellitus. *PLoS One*. 4(9):e7109](http://paperpile.com/b/a3wH5Q/EgzaA)

[Paun A, Yau C, Danska JS. 2016. Immune recognition and response to the intestinal microbiome in type 1 diabetes. *J. Autoimmun.* 71:10–18](http://paperpile.com/b/a3wH5Q/bskJY)

[Pereira GRC, Da Silva ANR, Do Nascimento SS, De Mesquita JF. 2019. In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. *J. Cell. Biochem.* 120(3):3583–98](http://paperpile.com/b/a3wH5Q/liFzT)

[Peterson LK, Fujinami RS. 2007. MOLECULAR MIMICRY](http://paperpile.com/b/a3wH5Q/3gtqM)

[Phelan J, Grabowska AD, Sepúlveda N. 2020. A potential antigenic mimicry between viral and human proteins linking Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) with autoimmunity: The case of HPV immunization. *Autoimmun. Rev.* 19(4):102487](http://paperpile.com/b/a3wH5Q/yv5LE)

[Polymeros D, Bogdanos DP, Day R, Arioli D, Vergani D, Forbes A. 2006. Does cross-reactivity between mycobacterium avium paratuberculosis and human intestinal antigens characterize Crohn’s disease? *Gastroenterology*. 131(1):85–96](http://paperpile.com/b/a3wH5Q/3tZqL)

[Polymeros D, Tsiamoulos ZP, Koutsoumpas AL, Smyk DS, Mytilinaiou MG, et al. 2014. Bioinformatic and immunological analysis reveals lack of support for measles virus related mimicry in Crohn’s disease. *BMC Med.* 12:139](http://paperpile.com/b/a3wH5Q/ay047)

[Poole BD, Scofield RH, Harley JB, James JA. 2006. Epstein-Barr virus and molecular mimicry in systemic lupus erythematosus. *Autoimmunity*. 39(1):63–70](http://paperpile.com/b/a3wH5Q/Ar7Lo)

[Proal AD, Marshall TG. 2018. Re-framing the theory of autoimmunity in the era of the microbiome: persistent pathogens, autoantibodies, and molecular mimicry. *Discov. Med.* 25(140):299–308](http://paperpile.com/b/a3wH5Q/2q89t)

[Qin L, Kufareva I, Holden LG, Wang C, Zheng Y, et al. 2015. Structural biology. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. *Science*. 347(6226):1117–22](http://paperpile.com/b/a3wH5Q/nzWN)

[Quayle AJ, Wilson KB, Li SG, Kjeldsen-Kragh J, Oftung F, et al. 1992. Peptide recognition, T cell receptor usage and HLA restriction elements of human heat-shock protein (hsp) 60 and mycobacterial 65-kDa hsp-reactive T cell clones from rheumatoid synovial fluid. *Eur. J. Immunol.* 22(5):1315–22](http://paperpile.com/b/a3wH5Q/Vi7Qz)

[Rahman SA, Schomburg D. 2006. Observing local and global properties of metabolic pathways: “load points” and “choke points” in the metabolic networks. *Bioinformatics*. 22(14):1767–74](http://paperpile.com/b/a3wH5Q/oCuTs)

[Rajagopalan K, Mooney SM, Parekh N, Getzenberg RH, Kulkarni P. 2011. A majority of the cancer/testis antigens are intrinsically disordered proteins. *J. Cell. Biochem.* 112(11):3256–67](http://paperpile.com/b/a3wH5Q/vnk3m)

[Raman K, Yeturu K, Chandra N. 2008. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Syst. Biol.* 2:109](http://paperpile.com/b/a3wH5Q/lNL7O)

[Rani PS, Sechi LA, Ahmed N. 2010. Mycobacterium avium subsp. paratuberculosis as a trigger of type-1 diabetes: destination Sardinia, or beyond? *Gut Pathog.* 2(1):1](http://paperpile.com/b/a3wH5Q/2MiS6)

[Rastogi N, Labrousse V, Goh KS. 1996. In vitro activities of fourteen antimicrobial agents against drug susceptible and resistant clinical isolates of Mycobacterium tuberculosis and comparative intracellular activities against the virulent H37Rv strain in human macrophages. *Curr. Microbiol.* 33(3):167–75](http://paperpile.com/b/a3wH5Q/Rm3cU)

[Reay PA, Kantor RM, Davis MM. 1994. Use of global amino acid replacements to define the requirements for MHC binding and T cell recognition of moth cytochrome c (93-103). *J. Immunol.* 152(8):3946–57](http://paperpile.com/b/a3wH5Q/Izr8G)

[Rey J, Rasolohery I, Tufféry P, Guyon F, Moroy G. 2019. PatchSearch: a web server for off-target protein identification. *Nucleic Acids Res.* 47(W1):W365–72](http://paperpile.com/b/a3wH5Q/0keoh)

[Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16(6):276–77](http://paperpile.com/b/a3wH5Q/1PgN6)

[Robertson RE, Cerf O, Condron RJ, Donaghy JA, Heggum C, Jordan K. 2017. Review of the controversy over whether or not Mycobacterium avium subsp. paratuberculosis poses a food safety risk with pasteurised dairy products](http://paperpile.com/b/a3wH5Q/rbgga)

[Rojas M, Restrepo-Jiménez P, Monsalve DM, Pacheco Y, Acosta-Ampudia Y, et al. 2018. Molecular mimicry and autoimmunity. *J. Autoimmun.* 95:100–123](http://paperpile.com/b/a3wH5Q/aYJP3)

[Rönnblom L, Alm GV. 2001. An etiopathogenic role for the type I IFN system in SLE. *Trends Immunol.* 22(8):427–31](http://paperpile.com/b/a3wH5Q/tbnpR)

[Rosen A, Casciola-Rosen L, Ahearn J. 1995. Novel packages of viral and self-antigens are generated during apoptosis. *J. Exp. Med.* 181(4):1557–61](http://paperpile.com/b/a3wH5Q/QbIF1)

[Rosenblum MD, Gratz IK, Paw JS, Abbas AK. 2012. Treating human autoimmunity: current practice and future prospects. *Sci. Transl. Med.* 4(125):125sr1](http://paperpile.com/b/a3wH5Q/92SP7)

[Rose NR. 2015. Molecular mimicry and clonal deletion: A fresh look. *J. Theor. Biol.* 375:71–76](http://paperpile.com/b/a3wH5Q/vX4ho)

[Rose NR, Mackay IR. 2000. Molecular mimicry: a critical look at exemplary instances in human diseases. *Cell. Mol. Life Sci.* 57(4):542–51](http://paperpile.com/b/a3wH5Q/AKFCg)

[Rothstein DM. 2016. Rifamycins, Alone and in Combination](http://paperpile.com/b/a3wH5Q/jAVNN)

[Ruan H, Sun Q, Zhang W, Liu Y, Lai L. 2019. Targeting intrinsically disordered proteins at the edge of chaos. *Drug Discov. Today*. 24(1):217–27](http://paperpile.com/b/a3wH5Q/J5PVT)

[Sámano-Sánchez H, Gibson TJ. 2020. Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity. *Trends Biochem. Sci.* 45(6):526–44](http://paperpile.com/b/a3wH5Q/Zs0m2)

[Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33(Web Server issue):W363–67](http://paperpile.com/b/a3wH5Q/1KSBE)

[Schulze K, Schultheiss HP. 1995. The role of the ADP/ATP carrier in the pathogenesis of viral heart disease. *Eur. Heart J.* 16 Suppl O:64–67](http://paperpile.com/b/a3wH5Q/Mw78O)

[Sechi LA, Dow CT. 2015. Mycobacterium avium ss. paratuberculosis Zoonosis - The Hundred Year War - Beyond Crohn’s Disease. *Front. Immunol.* 6:96](http://paperpile.com/b/a3wH5Q/xtUd5)

[Sechi LA, Gazouli M, Ikonomopoulos J, Lukas JC, Scanu AM, et al. 2005a. Mycobacterium avium subsp. paratuberculosis, genetic susceptibility to Crohn’s disease, and Sardinians: the way ahead. *J. Clin. Microbiol.* 43(10):5275–77](http://paperpile.com/b/a3wH5Q/iSKfT)

[Sechi L-A, Gazouli M, Sieswerda L-E, Molicotti P, Ahmed N, et al. 2006. Relationship between Crohn’s disease, infection with Mycobacterium avium subspecies paratuberculosis and SLC11A1 gene polymorphisms in Sardinian patients. *World J. Gastroenterol.* 12(44):7161–64](http://paperpile.com/b/a3wH5Q/R70go)

[Sechi LA, Scanu AM, Molicotti P, Cannas S, Mura M, et al. 2005b. Detection and Isolation of Mycobacterium avium subspecies paratuberculosis from intestinal mucosal biopsies of patients with and without Crohn’s disease in Sardinia. *Am. J. Gastroenterol.* 100(7):1529–36](http://paperpile.com/b/a3wH5Q/umP2Y)

[Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, et al. 2008. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36(Database issue):D351–59](http://paperpile.com/b/a3wH5Q/188CV)

[Sfriso P, Ghirardello A, Botsios C, Tonon M, Zen M, et al. 2010. Infections and autoimmunity: the multifaceted relationship. *J. Leukoc. Biol.* 87(3):385–95](http://paperpile.com/b/a3wH5Q/Iuig8)

[Sharma A, Pan A. 2012. Identification of potential drug targets in Yersinia pestis using metabolic pathway analysis: MurE ligase as a case study. *Eur. J. Med. Chem.* 57:185–95](http://paperpile.com/b/a3wH5Q/aHc8P)

[Sharp RC, Beg SA, Naser SA. 2018. Polymorphisms in Protein Tyrosine Phosphatase Non-receptor Type 2 and 22 (PTPN2/22) Are Linked to Hyper-Proliferative T-Cells and Susceptibility to Mycobacteria in Rheumatoid Arthritis](http://paperpile.com/b/a3wH5Q/GG2GH)

[Sheikh KA, Ho TW, Nachamkin I, Li CY, Cornblath DR, et al. 1998. Molecular mimicry in Guillain-Barré syndrome. *Ann. N. Y. Acad. Sci.* 845:307–21](http://paperpile.com/b/a3wH5Q/n3ktS)

[Sherman DR, Mdluli K, Hickey MJ, Arain TM, Morris SL, et al. 1996. Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis. *Science*. 272(5268):1641–43](http://paperpile.com/b/a3wH5Q/ipn2x)

[Shin SJ, Lee BS, Koh W-J, Manning EJB, Anklam K, et al. 2010. Efficient differentiation of Mycobacterium avium complex species and subspecies by use of five-target multiplex PCR. *J. Clin. Microbiol.* 48(11):4057–62](http://paperpile.com/b/a3wH5Q/gePeU)

[Shulman-Peleg A, Nussinov R, Wolfson HJ. 2005. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* 33(Web Server issue):W337–41](http://paperpile.com/b/a3wH5Q/ZCOko)

[Singh I, Yadav AR, Mohanty KK, Katoch K, Sharma P, et al. 2015. Molecular mimicry between Mycobacterium leprae proteins (50S ribosomal protein L2 and Lysyl-tRNA synthetase) and myelin basic protein: a possible mechanism of nerve damage in leprosy. *Microbes Infect.* 17(4):247–57](http://paperpile.com/b/a3wH5Q/pKnGV)

[Singh R, Wiseman B, Deemagarn T, Jha V, Switala J, Loewen PC. 2008. Comparative study of catalase-peroxidases (KatGs). *Arch. Biochem. Biophys.* 471(2):207–14](http://paperpile.com/b/a3wH5Q/tIIXI)

[Singh V, Mizrahi V. 2017. Identification and validation of novel drug targets in Mycobacterium tuberculosis. *Drug Discov. Today*. 22(3):503–9](http://paperpile.com/b/a3wH5Q/mYrAI)

[Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R](http://paperpile.com/b/a3wH5Q/GPil0)

[Sinigaglia F, Hammer J. 1994. Rules for peptide binding to MHC class II molecules. *APMIS*. 102(4):241–48](http://paperpile.com/b/a3wH5Q/wSgSo)

[Sisto M, Cucci L, D’Amore M, Dow TC, Mitolo V, Lisi S. 2010. Proposing a relationship between Mycobacterium avium subspecies paratuberculosis infection and Hashimoto’s thyroiditis](http://paperpile.com/b/a3wH5Q/ToAtc)

[Siwo GH, Tan A, Button-Simons KA, Samarakoon U, Checkley LA, et al. 2015. Predicting functional and regulatory divergence of a drug resistance transporter gene in the human malaria parasite. *BMC Genomics*. 16:115](http://paperpile.com/b/a3wH5Q/KuHvh)

[Smatti MK, Cyprian FS, Nasrallah GK, Al Thani AA, Almishal RO, Yassine HM. 2019. Viruses and Autoimmunity: A Review on the Potential Interaction and Molecular Mechanisms. *Viruses*. 11(8):](http://paperpile.com/b/a3wH5Q/mgUDU)

[Society RC of TBT, Research Committee of the British Thoracic Society. 2001. First randomised trial of treatments for pulmonary disease caused by M avium intracellulare, M malmoense, and M xenopi in HIV negative patients: rifampicin, ethambutol and isoniazid versus rifampicin and ethambutol](http://paperpile.com/b/a3wH5Q/Lgeht)

[Spriggs CC, Harwood MC, Tsai B. 2019. How non-enveloped viruses hijack host machineries to cause infection](http://paperpile.com/b/a3wH5Q/eIyTo)

[Stebbins CE, Galán JE. 2001. Structural mimicry in bacterial virulence. *Nature*. 412(6848):701–5](http://paperpile.com/b/a3wH5Q/k3Ra)

[Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45(D1):D362–68](http://paperpile.com/b/a3wH5Q/5Kkbq)

[Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. 2016. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data](http://paperpile.com/b/a3wH5Q/yq8X7)

[Tahir M, Hayat M, Khan SA. 2018. A Two-Layer Computational Model for Discrimination of Enhancer and Their Types Using Hybrid Features Pace of Pseudo K-Tuple Nucleotide Composition](http://paperpile.com/b/a3wH5Q/ilyca)

[Talbot PJ, Paquette JS, Ciurli C, Antel JP, Ouellet F. 1996. Myelin basic protein and human coronavirus 229E cross-reactive T cells in multiple sclerosis. *Ann. Neurol.* 39(2):233–40](http://paperpile.com/b/a3wH5Q/6E11U)

[Tarakhovsky A, Prinjha RK. 2018. Drawing on disorder: How viruses use histone mimicry to their advantage. *J. Exp. Med.* 215(7):1777–87](http://paperpile.com/b/a3wH5Q/P4ZQt)

[Tha-In T, Bayry J, Metselaar HJ, Kaveri SV, Kwekkeboom J. 2008. Modulation of the cellular immune system by intravenous immunoglobulin. *Trends Immunol.* 29(12):608–15](http://paperpile.com/b/a3wH5Q/Ah49f)

[Thaper D, Prabha V. 2018. Molecular mimicry: An explanation for autoimmune diseases and infertility. *Scand. J. Immunol.* 88(2):e12697](http://paperpile.com/b/a3wH5Q/SWLY9)

[Tian L, Chou H-L, Fukuda M, Kumamaru T, Okita TW. 2020. mRNA Localization in Plant Cells. *Plant Physiol.* 182(1):97–109](http://paperpile.com/b/a3wH5Q/FlMjU)

[Tiwana H, Wilson C, Alvarez A, Abuknesha R, Bansal S, Ebringer A. 1999. Cross-reactivity between the rheumatoid arthritis-associated motif EQKRAA and structurally related sequences found in Proteus mirabilis. *Infect. Immun.* 67(6):2769–75](http://paperpile.com/b/a3wH5Q/jIiK5)

[Vandepoele K, Simillion C, Van de Peer Y. 2002. Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.* 18(12):606–8](http://paperpile.com/b/a3wH5Q/VwBzc)

[van der Kooij SM, de Vries-Bouwstra JK, Goekoop-Ruiterman YPM, van Zeben D, Kerstens PJSM, et al. 2007. Limited efficacy of conventional DMARDs after initial methotrexate failure in patients with recent onset rheumatoid arthritis treated according to the disease activity score. *Ann. Rheum. Dis.* 66(10):1356–62](http://paperpile.com/b/a3wH5Q/37qW9)

[Van de Water J, Turchany J, Leung PS, Lake J, Munoz S, et al. 1993. Molecular mimicry in primary biliary cirrhosis. Evidence for biliary epithelial expression of a molecule cross-reactive with pyruvate dehydrogenase complex-E2. *J. Clin. Invest.* 91(6):2653–64](http://paperpile.com/b/a3wH5Q/uc4oX)

[Vatti A, Monsalve DM, Pacheco Y, Chang C, Anaya J-M, Gershwin ME. 2017. Original antigenic sin: A comprehensive review. *J. Autoimmun.* 83:12–21](http://paperpile.com/b/a3wH5Q/RP1Xg)

[Vo AA, Lukovsky M, Toyoda M, Wang J, Reinsmoen NL, et al. 2008. Rituximab and intravenous immune globulin for desensitization during renal transplantation. *N. Engl. J. Med.* 359(3):242–51](http://paperpile.com/b/a3wH5Q/ftAOn)

[Wallace AC, Laskowski RA, Thornton JM. 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions](http://paperpile.com/b/a3wH5Q/MwPJw)

[Wang H, Yuan F-F, Dai Z-W, Wang B, Ye D-Q. 2018. Association between rheumatoid arthritis and genetic variants of natural resistance-associated macrophage protein 1 gene: A meta-analysis. *Int. J. Rheum. Dis.* 21(9):1651–58](http://paperpile.com/b/a3wH5Q/JkYvs)

[*Website*.](http://paperpile.com/b/a3wH5Q/o713m) [ttps://github.com/owlcollab/owltools](about:blank)

[Weis BL, Schleiff E, Zerges W. 2013. Protein targeting to subcellular organelles via MRNA localization. *Biochim. Biophys. Acta*. 1833(2):260–73](http://paperpile.com/b/a3wH5Q/JSP4h)

[Wekerle H, Hohlfeld R. 2003. Molecular Mimicry in Multiple Sclerosis](http://paperpile.com/b/a3wH5Q/gDy9t)

[Wengenack NL, Jensen MP, Rusnak F, Stern MK. 1999. Mycobacterium tuberculosis KatG is a peroxynitritase. *Biochem. Biophys. Res. Commun.* 256(3):485–87](http://paperpile.com/b/a3wH5Q/eazSA)

[Whiley H, Keegan A, Giglio S, Bentham R. 2012. Mycobacterium avium complex--the role of potable water in disease transmission. *J. Appl. Microbiol.* 113(2):223–32](http://paperpile.com/b/a3wH5Q/UePoy)

[Wildner G, Diedrichs-Möhring M. 2003. Autoimmune uveitis induced by molecular mimicry of peptides from rotavirus, bovine casein and retinal S-antigen. *Eur. J. Immunol.* 33(9):2577–87](http://paperpile.com/b/a3wH5Q/kiahH)

[Winthrop KL, Chiller T. 2009. Preventing and treating biologic-associated opportunistic infections. *Nat. Rev. Rheumatol.* 5(7):405–10](http://paperpile.com/b/a3wH5Q/SK9hy)

[Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018](http://paperpile.com/b/a3wH5Q/loEpU)

[Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36(Database issue):D901–6](http://paperpile.com/b/a3wH5Q/IRZIf)

[Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases](http://paperpile.com/b/a3wH5Q/a9Jy7)

[Wright PE, Dyson HJ. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16(1):18–29](http://paperpile.com/b/a3wH5Q/cRphE)

[Wucherpfennig KW. 2001. Mechanisms for the induction of autoimmunity by infectious agents. *J. Clin. Invest.* 108(8):1097–1104](http://paperpile.com/b/a3wH5Q/K19KO)

[Wucherpfennig KW, Allen PM, Celada F, Cohen IR, De Boer R, et al. 2007. Polyspecificity of T cell and B cell receptor recognition. *Semin. Immunol.* 19(4):216–24](http://paperpile.com/b/a3wH5Q/2N5Si)

[Wucherpfennig KW, Sette A, Southwood S, Oseroff C, Matsui M, et al. 1994. Structural requirements for binding of an immunodominant myelin basic protein peptide to DR2 isotypes and for its recognition by human T cell clones. *J. Exp. Med.* 179(1):279–90](http://paperpile.com/b/a3wH5Q/6KL82)

[Wucherpfennig KW, Strominger JL. 1995. Molecular mimicry in T cell-mediated autoimmunity: Viral peptides activate human T cell clones specific for myelin basic protein](http://paperpile.com/b/a3wH5Q/lu99O)

[Wu H, Deng Y, Feng Y, Long D, Ma K, et al. 2018. Epigenetic regulation in B-cell maturation and its dysregulation in autoimmunity. *Cell. Mol. Immunol.* 15(7):676–84](http://paperpile.com/b/a3wH5Q/dJ2oD)

[Wynne JW, Bull TJ, Seemann T, Bulach DM, Wagner J, et al. 2011. Exploring the zoonotic potential of Mycobacterium avium subspecies paratuberculosis through comparative genomics. *PLoS One*. 6(7):e22171](http://paperpile.com/b/a3wH5Q/6ZKR5)

[Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, et al. 2014. Structural disorder in viral proteins. *Chem. Rev.* 114(13):6880–6911](http://paperpile.com/b/a3wH5Q/dwAvu)

[Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. 2010. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*. 1804(4):996–1010](http://paperpile.com/b/a3wH5Q/ZXMuz)

[Xue B, Dunker AK, Uversky VN. 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* 30(2):137–49](http://paperpile.com/b/a3wH5Q/jOCCN)

[Xue B, Uversky VN. 2014. Intrinsic Disorder in Proteins Involved in the Innate Antiviral Immunity: Another Flexible Side of a Molecular Arms Race. *J. Mol. Biol.* 426(6):1322–50](http://paperpile.com/b/a3wH5Q/apqvA)

[Yamada K, Gherasim C, Banerjee R, Koutmos M. 2015. Structure of Human B12 Trafficking Protein CblD Reveals Molecular Mimicry and Identifies a New Subfamily of Nitro-FMN Reductases. *J. Biol. Chem.* 290(49):29155–66](http://paperpile.com/b/a3wH5Q/miad)

[Yang J, Gao M, Xiong J, Su Z, Huang Y. 2019. Features of molecular recognition of intrinsically disordered proteins via coupled folding and binding. *Protein Sci.* 28(11):1952–65](http://paperpile.com/b/a3wH5Q/HPsVt)

[Yan Z, Lécuyer E, Blanchette M. 2019. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*. 35(14):i333–42](http://paperpile.com/b/a3wH5Q/y0tdN)

[Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. 2004. Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* 14(5):917–24](http://paperpile.com/b/a3wH5Q/hXqnX)

[Yusung S, Braun J. 2014. Molecular mimicry, inflammatory bowel disease, and the vaccine safety debate. *BMC Med.* 12:166](http://paperpile.com/b/a3wH5Q/F7KG9)

[Zabriskie JB, Freimer EH. 1966. AN IMMUNOLOGICAL RELATIONSHIP BETWEEN THE GROUP A STREPTOCOCCUS AND MAMMALIAN MUSCLE](http://paperpile.com/b/a3wH5Q/j0QVL)

[Zhang R, Ou H-Y, Zhang C-T. 2004. DEG: a database of essential genes. *Nucleic Acids Res.* 32(Database issue):D271–72](http://paperpile.com/b/a3wH5Q/x1jOF)

[Zhang T, Tan P, Wang L, Jin N, Li Y, et al. 2017. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.* 45(D1):D135–38](http://paperpile.com/b/a3wH5Q/d7fbS)

[Zhu J, Paul WE. 2008. CD4 T cells: fates, functions, and faults. *Blood*. 112(5):1557–69](http://paperpile.com/b/a3wH5Q/5YUCi)

[Zhu L, Qin J. 2019. A Viral Protein Mimics Histone to Hijack Host MORC3](http://paperpile.com/b/a3wH5Q/FMQox)

[Zur Hausen H. 2009. The search for infectious causes of human cancers: where and why. *Virology*. 392(1):1–10](http://paperpile.com/b/a3wH5Q/RBMO)