

Knowledge-based systems for assessing immunogenicity

Vladimir Brusnic, PhD MBA

Director of Bioinformatics

Cancer Vaccine Center

Dana-Farber Cancer Institute

SUMMARY

Introduction

Definitions

Framework

Immunogenicity

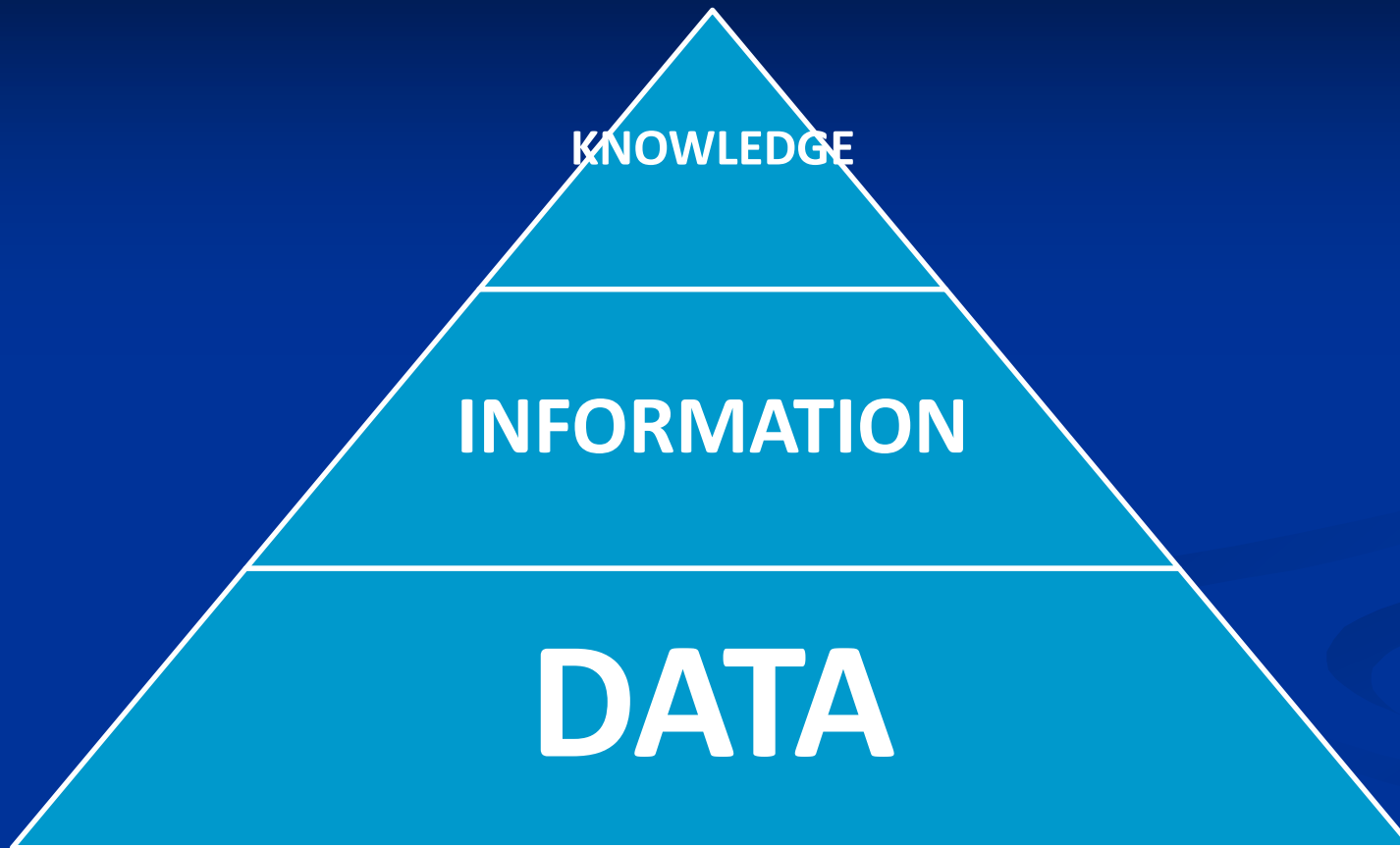
Examples

Malaria

FLAVIdB

Conclusions

Traditional databases contain:



Focus on extraction of information. Knowledge extraction and acquisition left to the user.

Protein Protein interferon gamma homo Search

[Save search](#) [Limits](#) [Advanced](#)[Help](#)[Display Settings:](#) Summary, 20 per page, Sorted by Default order[Send to:](#)

Filter your results:

Results: 1 to 20 of 642

<< First < Prev Page 1 of 33 Next > Last >>

 [interferon-gamma \[Homo sapiens\]](#)

1. 166 aa protein

Accession: AAB59534.1 GI: 184639

[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)[Manage Filters](#) [interferon-gamma \[Homo sapiens\]](#)

2. 144 aa protein

Accession: AAM28885.1 GI: 20805896

[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [interferon-gamma \[Homo sapiens\]](#)

3. 249 aa protein

Accession: AAA53230.1 GI: 551491

[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#) [interferon-gamma \[Homo sapiens\]](#)

4. 249 aa protein

Accession: AAA16521.1 GI: 186513

[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#) [interferon gamma \[Homo sapiens\]](#)

5. 143 aa protein

Accession: AAK95388.1 GI: 15421195

[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [interferon gamma \[Homo sapiens\]](#)

All (642)

[Bacteria \(2\)](#)[Related Structures \(558\)](#)[RefSeq \(161\)](#)▼ Top Organisms [\[Tree\]](#)

Homo sapiens (587)

synthetic construct (19)

Mus musculus (8)

Monkeypox virus (6)

Cowpox virus (5)

All other taxa (17)

[More...](#)

Find related data

Database:

Search details

interferon gamma[All Fields]
AND ("Homo"[Organism] OR

Protein Protein Search

Limits Advanced

Help

Display Settings: GenPept

Send to:

Change region shown

Customize view

interferon-gamma [Homo sapiens]

GenBank: AAB59534.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS AAB59534 166 aa linear PRI 17-APR-2002

DEFINITION interferon-gamma [Homo sapiens].

ACCESSION AAB59534

VERSION AAB59534.1 GI:184639

DBSOURCE locus HUMIFNG accession [J00219.1](#)

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 3 (residues 1 to 166)

AUTHORS Gray, P.W. and Goeddel, D.V.

TITLE Structure of the human immune interferon gene

JOURNAL Nature 298 (5877), 859-863 (1982)

PUBMED [6180322](#)

REFERENCE 7 (residues 1 to 166)

AUTHORS Gray, P.W. and Goeddel, D.V.

TITLE Human immune interferon (IFN-gamma) gene sequence and structure

JOURNAL Basic Life Sci. 25, 35-61 (1983)

PUBMED [6305337](#)

COMMENT Immune interferon (IFN-gamma) shows little homology with leukocyte (IFN-alpha) and fibroblast (IFN-beta) interferons. See other loci beginning <HUMIFNG>. IFN-gamma protein (35,000 to 70,000 mw) as

Analyze this sequence

[Run BLAST](#)[Identify Conserved Domains](#)[Find in this Sequence](#)

Articles about the IFNG gene

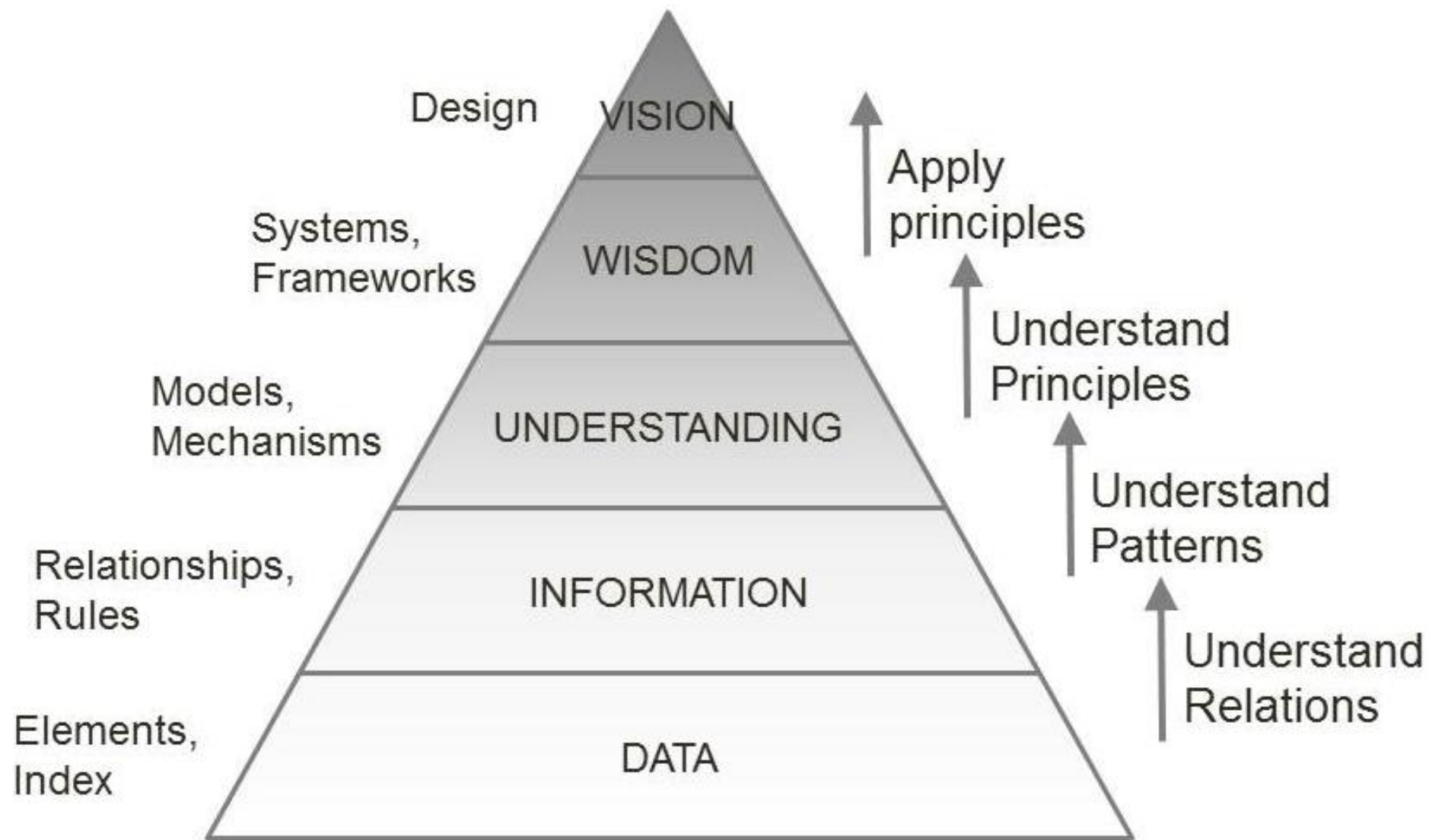
[Factors associated with in vitro interferon-gamma production in tubercul \[J Formos Med Assoc. 2011\]](#)[Enhanced demethylation of interferon-gamma promoter in peripheral bl \[Tohoku J Exp Med. 2011\]](#)[Functional polymorphism of interferon-gamma \(IFN-gamma\) gene +874T/A polymorphi: \[Prague Med Rep. 2011\]](#)

See all...

Identical proteins for AAB59534.1

[Sequence 2 from patent US 7964400 \[AEJ750271\]](#)[Sequence 20 from patent US 7795210 \[ADS432701\]](#)[Sequence 212 from patent US 774066 \[ADR870191\]](#)

See all...



The D2V knowledge hierarchy model. The layers from data to vision decrease in quantity of the output but increase in the overall value of knowledge, acquisition cost, and complexity of management. Lower layers are more amenable to algorithmic approach and automation, while higher levels increase in utilization of conceptualization and modeling approaches.

| LEVEL | INFORMATION SCIENCE | BASIC SCIENCE | APPLIED SCIENCE | CLINICAL SCIENCE | ENGINEERING/ TECHNOLOGY |
|---------------|------------------------|-------------------------|------------------|-------------------|-------------------------|
| Data | data | experiment | model elements | test, observation | product elements |
| Information | information | statistics | model evaluation | procedure | specifications |
| Understanding | <i>in silico</i> model | description (structure) | validated model | diagnosis | design |
| Wisdom | simulation | description (mechanism) | model deployment | treatment | prototype |
| Vision | prediction, decision | model, theory | design | design | product |

D2V hierarchy across disciplines

| LEVEL | INFORMATION SCIENCE | BASIC SCIENCE | APPLIED SCIENCE | CLINICAL SCIENCE | ENGINEERING/ TECHNOLOGY |
|---------------|------------------------|-------------------------|------------------|-------------------|-------------------------|
| Data | data | experiment | model elements | test, observation | product elements |
| Information | information | statistics | model evaluation | procedure | specifications |
| Understanding | <i>in silico</i> model | description (structure) | validated model | diagnosis | design |
| Wisdom | simulation | description (mechanism) | model deployment | treatment | prototype |
| Vision | prediction, decision | model, theory | design | design | product |

The diagram illustrates the D2V hierarchy across disciplines. Red arrows indicate the following relationships:

- Solid red arrows point from 'product elements' to 'test, observation', 'test, observation' to 'model elements', 'model elements' to 'experiment', 'experiment' to 'data', 'product elements' to 'specifications', 'specifications' to 'design', 'design' to 'prototype', 'prototype' to 'product', 'model elements' to 'validated model', 'validated model' to 'model deployment', 'model deployment' to 'design', 'description (mechanism)' to 'model, theory', and 'description (structure)' to 'model, theory'.
- Dashed red arrows point from 'test, observation' to 'procedure', 'procedure' to 'diagnosis', 'diagnosis' to 'treatment', 'treatment' to 'design', 'design' to 'product', 'validated model' to 'diagnosis', 'model deployment' to 'treatment', and 'design' to 'product'.

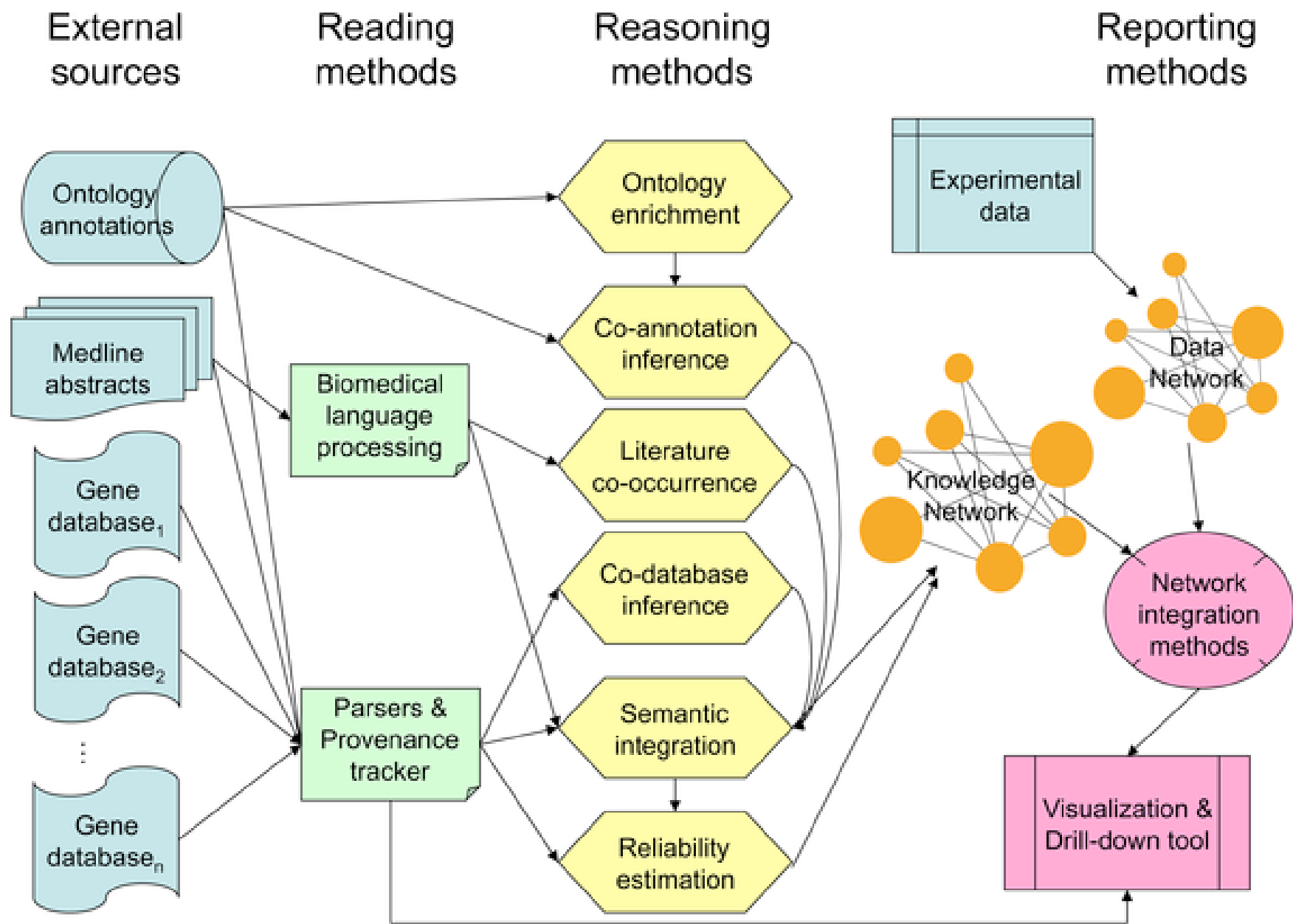
D2V hierarchy across disciplines

Knowledge based systems (KBS) are:

Tools of Artificial Intelligence applied to a narrow domain to provide analysis, decisions, and explanation. KBS are used for knowledge extraction and decision support.

KBS use:

- knowledge representation to document knowledge
- conceptual modeling to capture system properties
- learning to acquire new knowledge
- inference mechanisms to produce reasoning and explanation



Immunogenicity

- The ability of a substance (antigen) to induce immune response
- Not inherent property, requires a partner (a T or B cell)
- Individual-, profile-, history-, challenge- specific

Immunogenicity, a poorly-defined term

- Different types of responses possible, fine tuning
- Every target is immunogenic, given the right context
- Different laboratory, functional, and clinical outcomes
- Differs between individuals

DIFFERENT OBJECTIVES

PROTEINS

- Undesirable or desirable immunogenicity

INFLUENZA

- Diversity of new variant targets

DENGUE

- Existing immunogenicity against serotypes

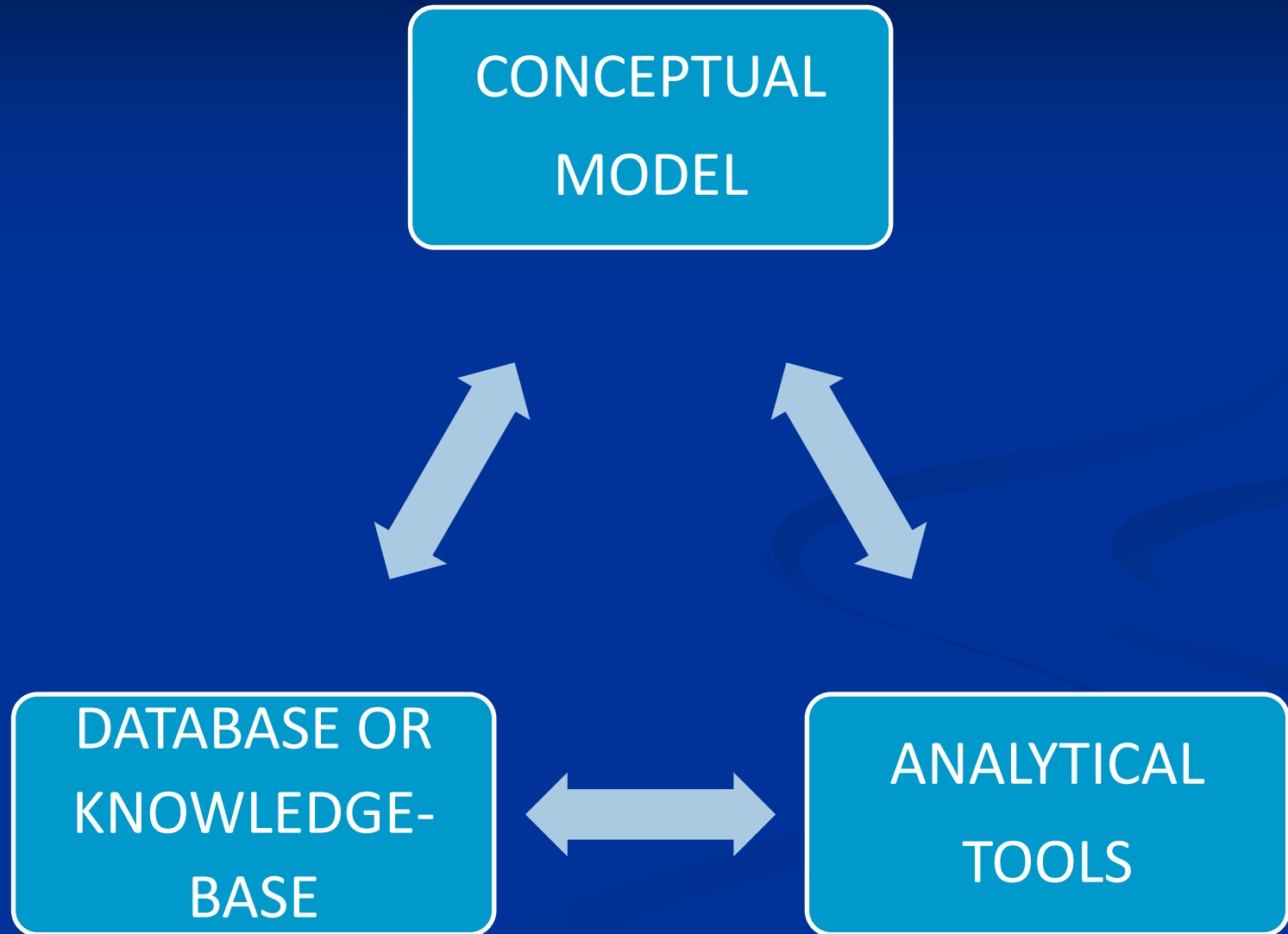
MALARIA

- Targeting large proteome

CANCER

- Targeting cancer-specific antigens

KBS COMPONENTS



SUMMARY

Introduction

Definitions

Framework

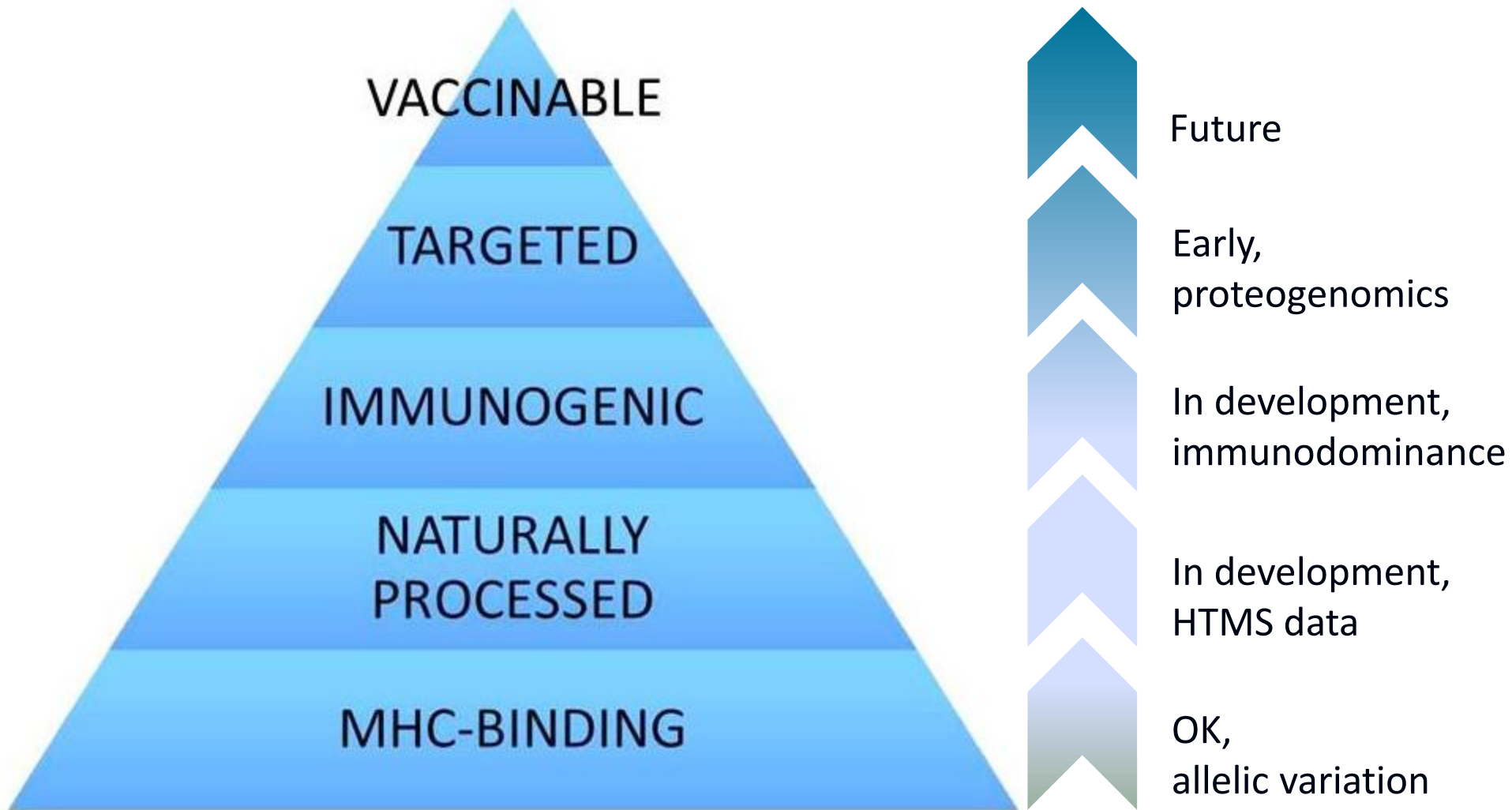
Examples

Malaria

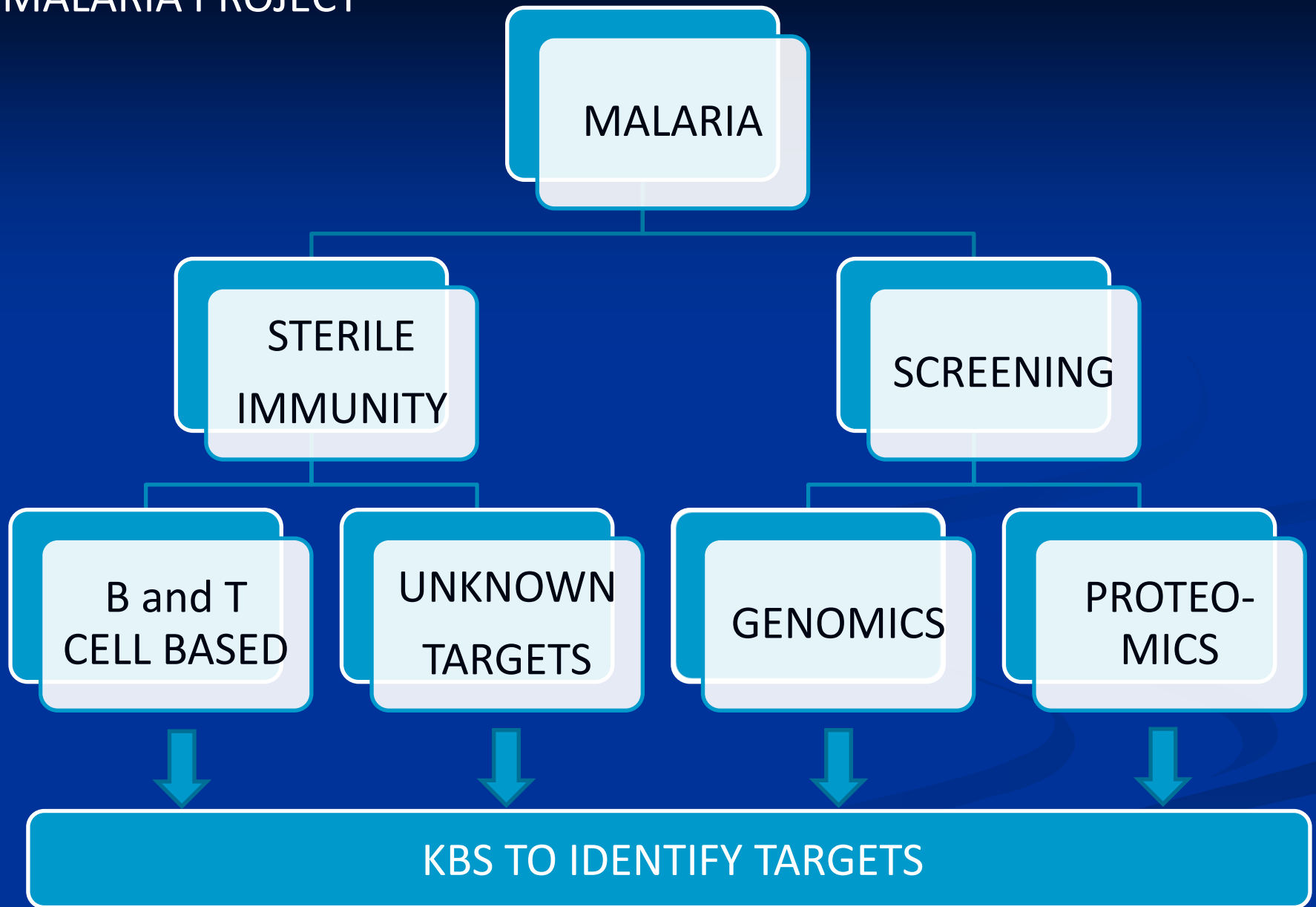
FLAVIdB

Conclusions

COMBINATORIALLY COMPLEX PROBLEM EPIOTOPE-BASED TARGET DISCOVERY FILTERS



MALARIA PROJECT



Preliminary data

19 proteins identified as pooled targets of human immune responses against *P. falciparum*. (Trieu et al., 2011)

958 transcripts identified in the gene expression study in *P. yoelii* (Mackellar et al, 2011).

Too many transcripts and too few proteins.

Need for a KBS for identification of vaccine targets.

Transcripts

- 958 targets; three time points; 623,155 9-mers



T-cell epitope prediction

- Neural network predictors: 100s to 1000s of targets



Proteomics

Naturally processed

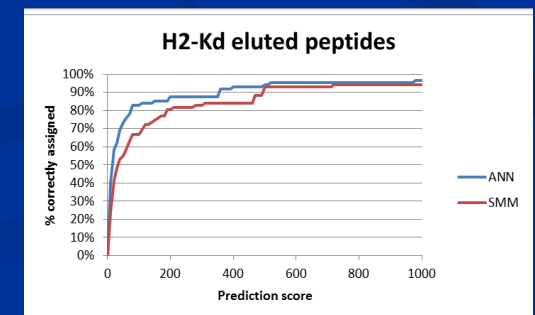
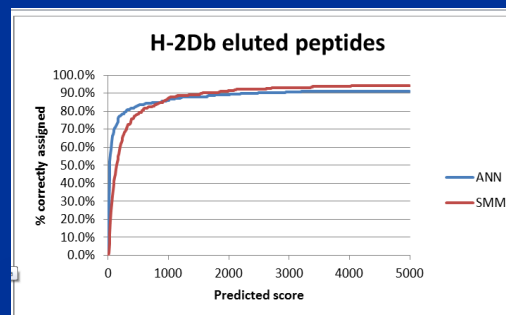
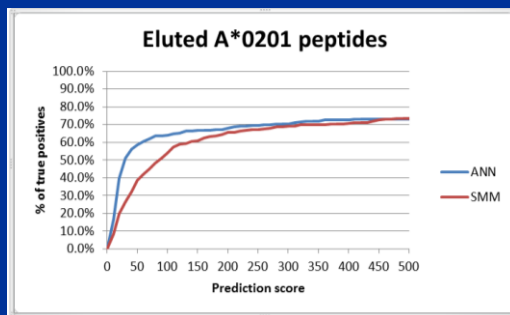
- Predictors of eluted motifs



KBS

Number of predicted peptides for YNN (24 hour liver infection)

| | Prediction threshold | | | | | | LENGTH | CAPTURE |
|-------------------------|----------------------|------|------|-----|-----|------|--------|---------|
| | 10 | 20 | 50 | 100 | 200 | 1000 | | |
| A0201 | 340 | 1202 | 2483 | | | | 9-mer | 63.40% |
| H-2D^d | | | 1 | 11 | 28 | 152 | 9-mer | 72.70% |
| H-2K^d | 304 | 534 | 870 | | | | 9-mer | 55.20% |
| H-2D^b | 74 | 169 | 331 | 510 | | | 9-mer | 69.40% |
| H-2K^b | 228 | 433 | 819 | | | | 8-mer | 77.90% |



SUMMARY

Introduction

Definitions

Framework

Examples

Malaria

FLAVIdB

Conclusions

How to convert data to application?

A next generation database:

Contains database, knowledgebase, analytical tools, and workflows.

Enables us to extract new knowledge from existing data, information, and knowledge base in automated fashion.

“Next generation databases”



Primary databases



Literature



Applied computing



Next generation database



Knowledge

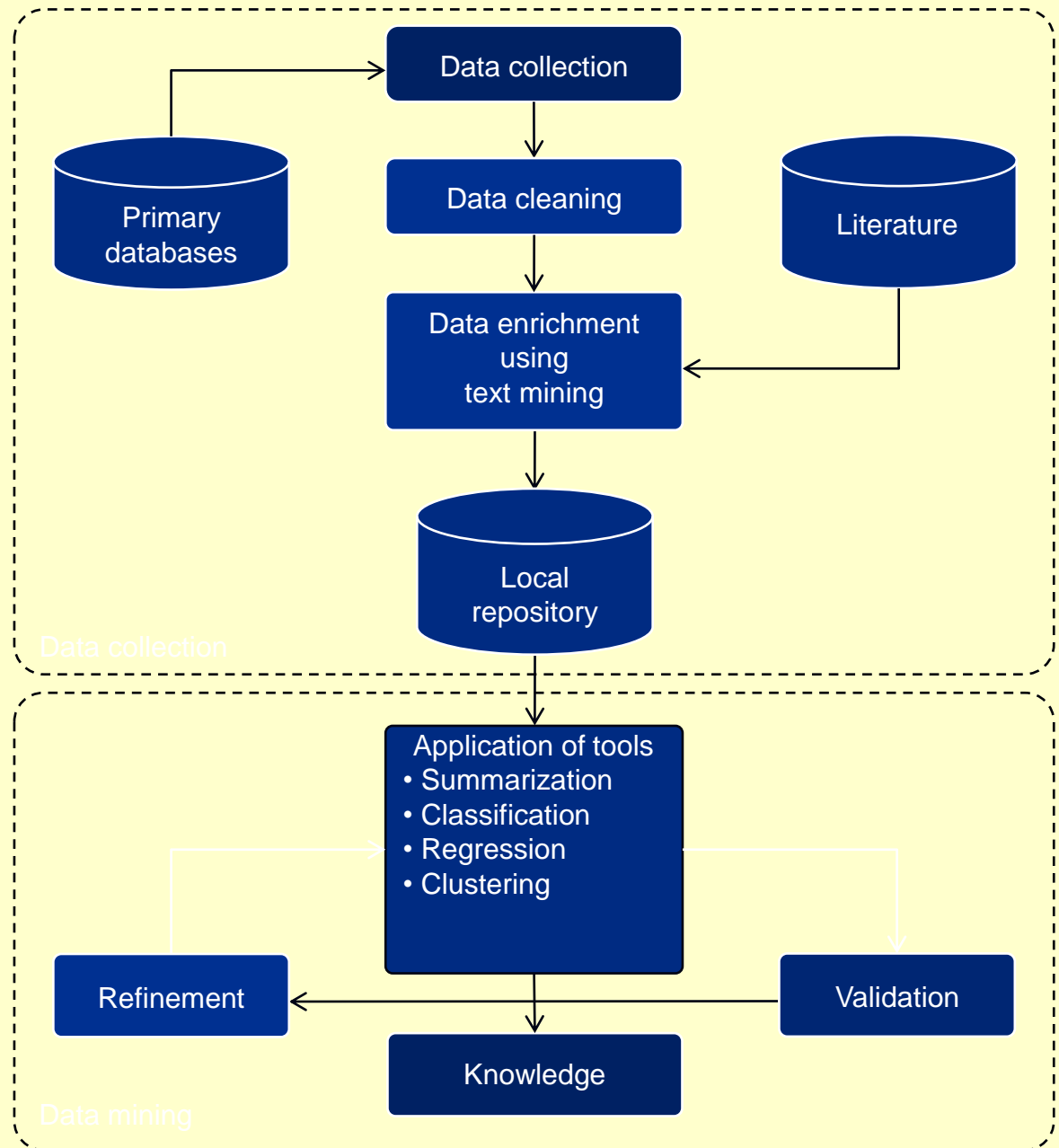


Experimental validation/
Clinical testing



Vaccine

FLAVIdB Architecture



FLAVIdb

- Of more than 70 species of flaviviruses approx. half causes human disease.
- Five most relevant human flaviviruses alone put half of the world's population at risk of infection.
- Flaviviruses have small single-stranded RNA genomes, ideal for rapid integration into a KBS.
- Dengue virus present an interesting challenge to vaccinology because of complications associated with secondary infection

The flavivirus proteome

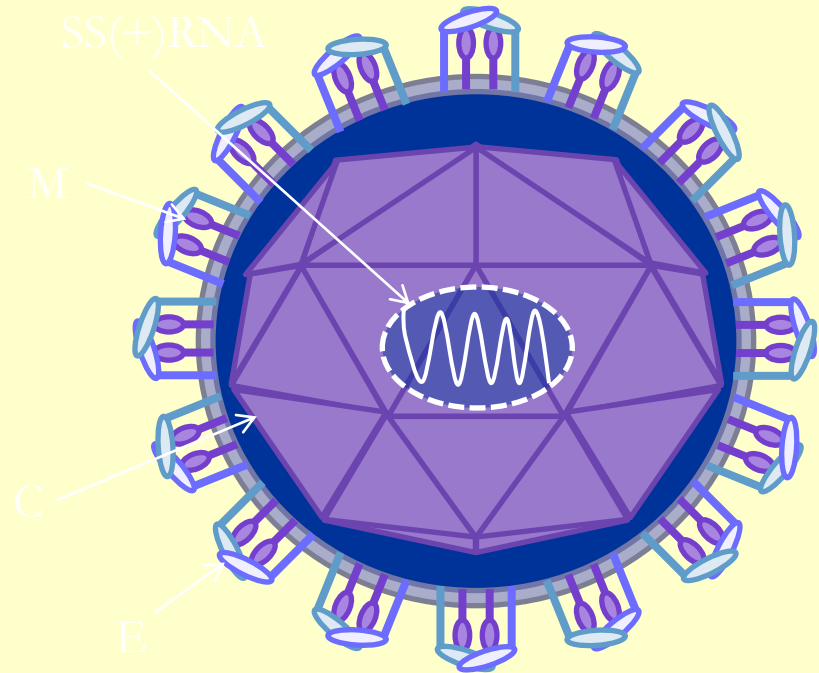
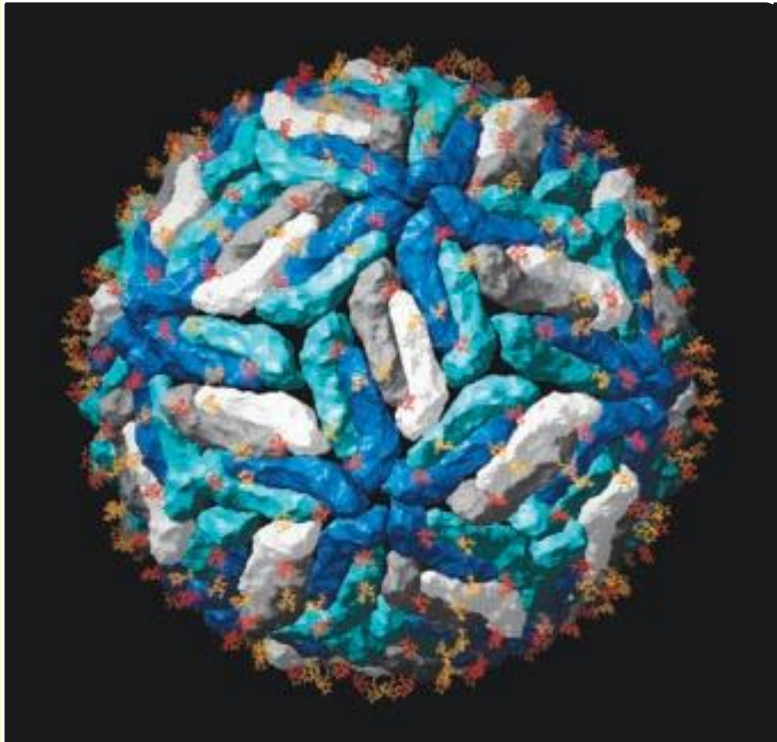
Flavivirus linear RNA proteome:

| | | | | | | | | | |
|-----|-----|---|-----|------|----|-----|----|------|-----|
| anC | prM | E | NS1 | NS2A | 2B | NS3 | 4A | NS4B | NS5 |
|-----|-----|---|-----|------|----|-----|----|------|-----|

Structural proteins

Non-structural proteins

Enveloped nucleocapsid - 50 nm in diameter.



Three classes of computational tools:

Data retrieval tools

- Keyword search
- BLAST query

Basic analytical tools

- Species classification
- Multiple sequence alignment
- Conservation and variability analysis
- Block entropy analysis

Advanced analytical tools

- Prediction of T-cell epitopes
- Analysis of shared features of neutralizing B-cell epitopes (BBscore)

Tools

Common approach to discovery of T-cell epitopes as vaccine targets

1. Selection of sequences of interest
2. Application of T-cell epitope prediction algorithms

Output: Epitope candidates (data + information)

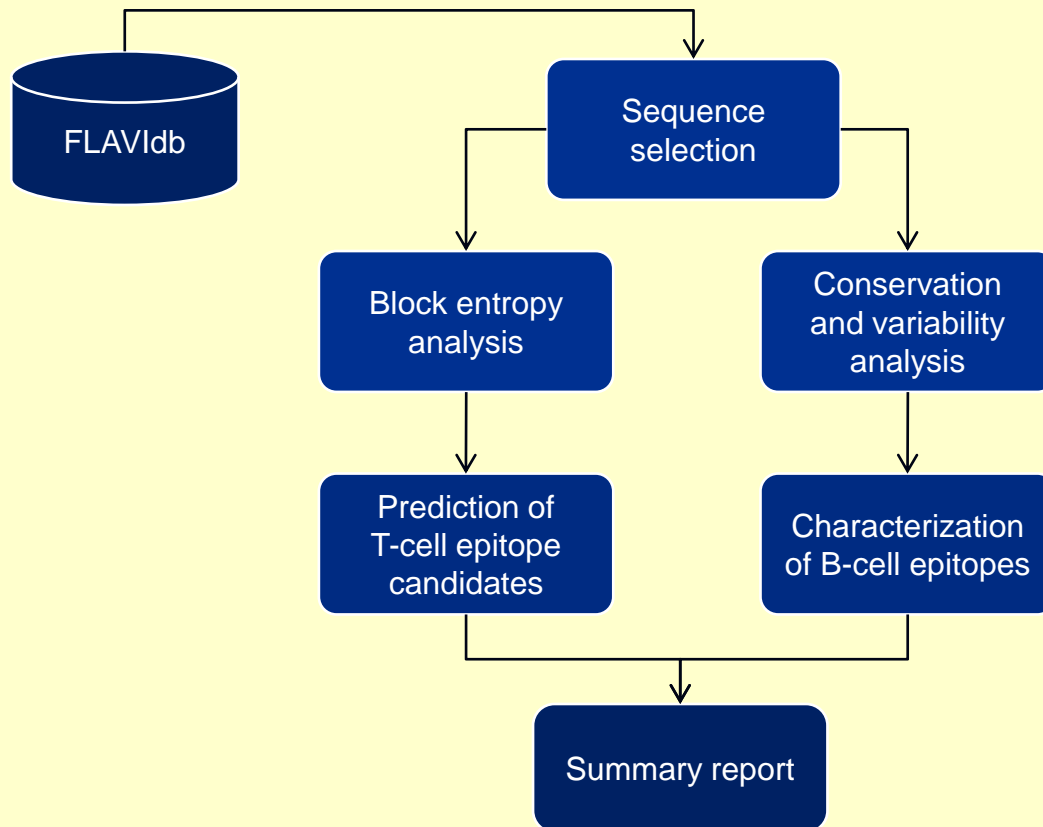
3. Conservation and variability analysis to ensure broad coverage of pathogen population and host population

Output: Reduced target set (data + information)

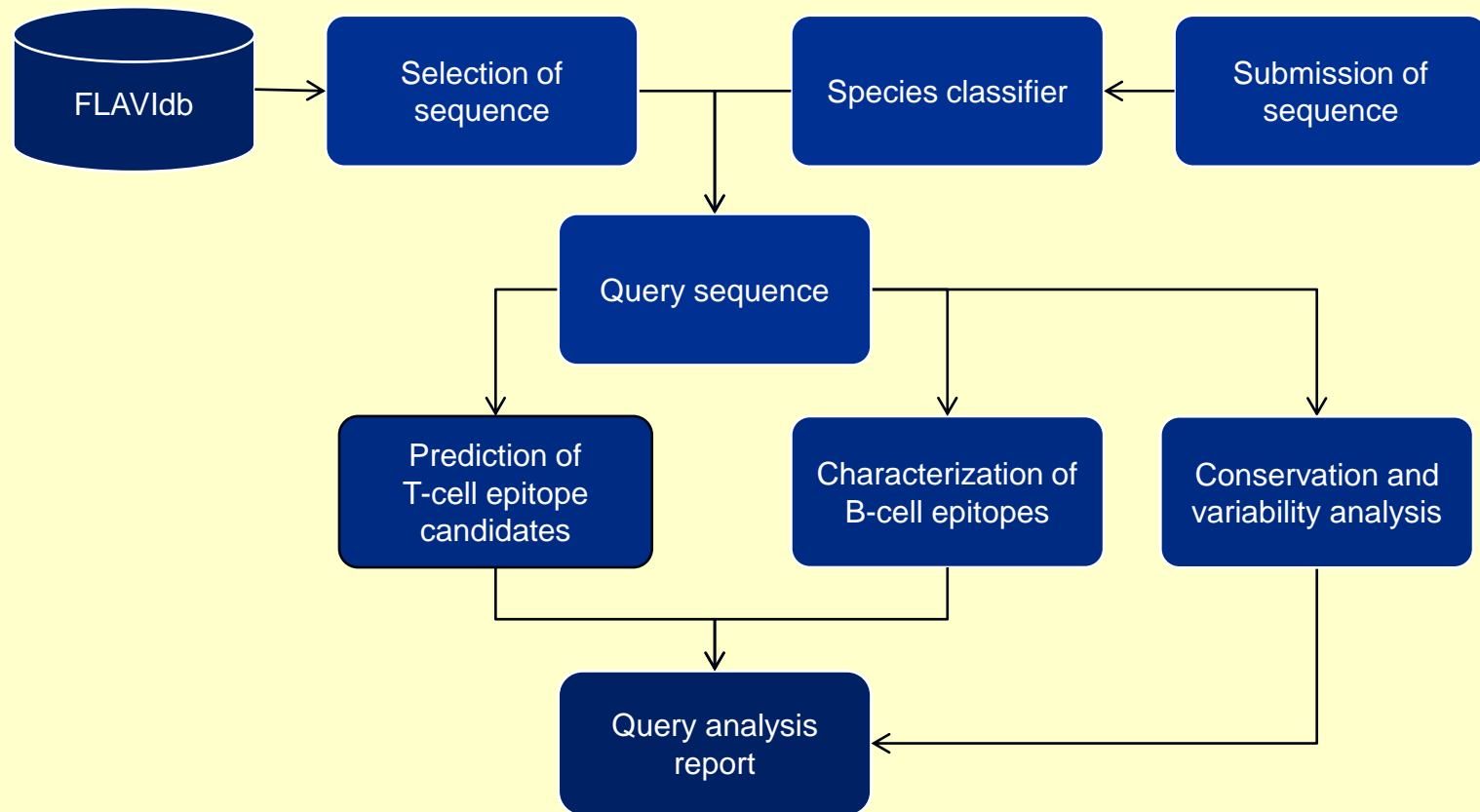
4. Experimental validation of epitope candidates

Output: Validated and characterized targets and their context (knowledge)

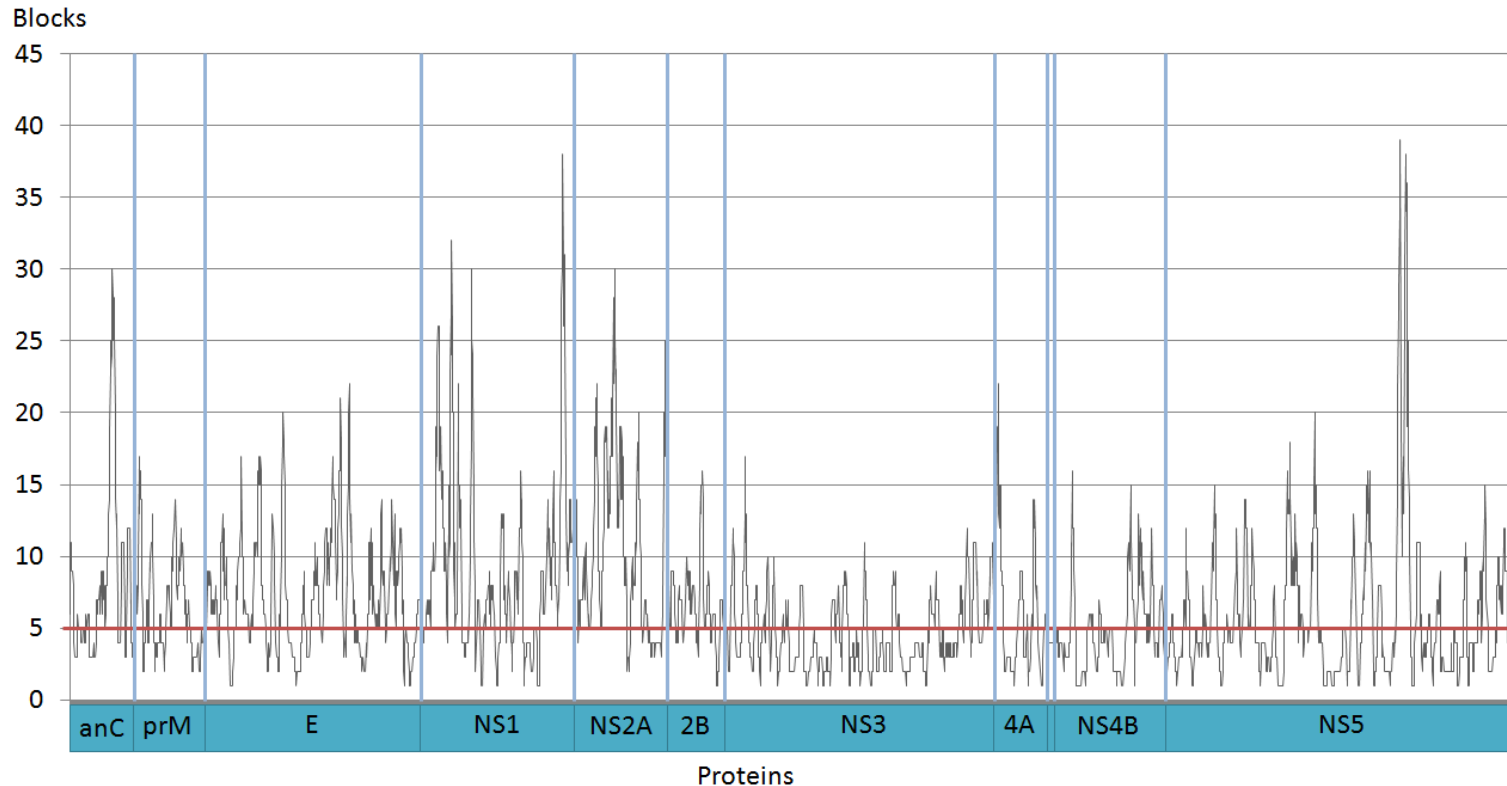
Workflows: Summary workflow



Workflows – Query analysis workflow



Blocks conservation analysis



1.551 blocks of 9-mers (5.113 peptides) cover 99% of DENV1-4 with 5 or less peptides

...As compared to 165 conserved 9-mers, by traditional conservation analysis (Khan *et al.*)

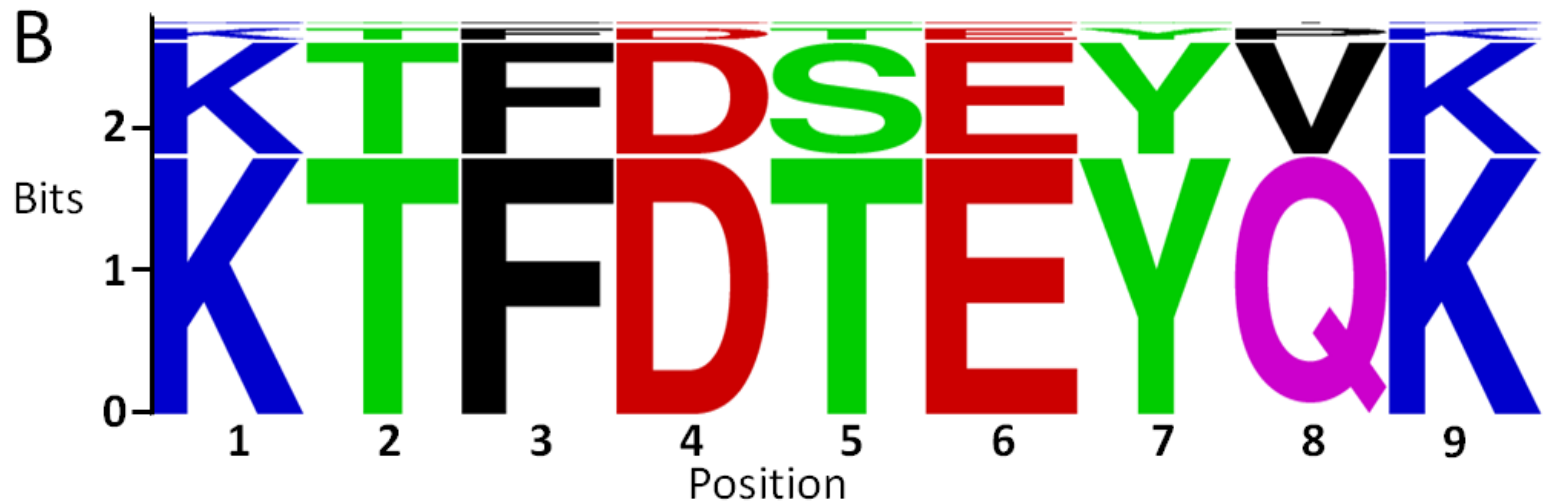
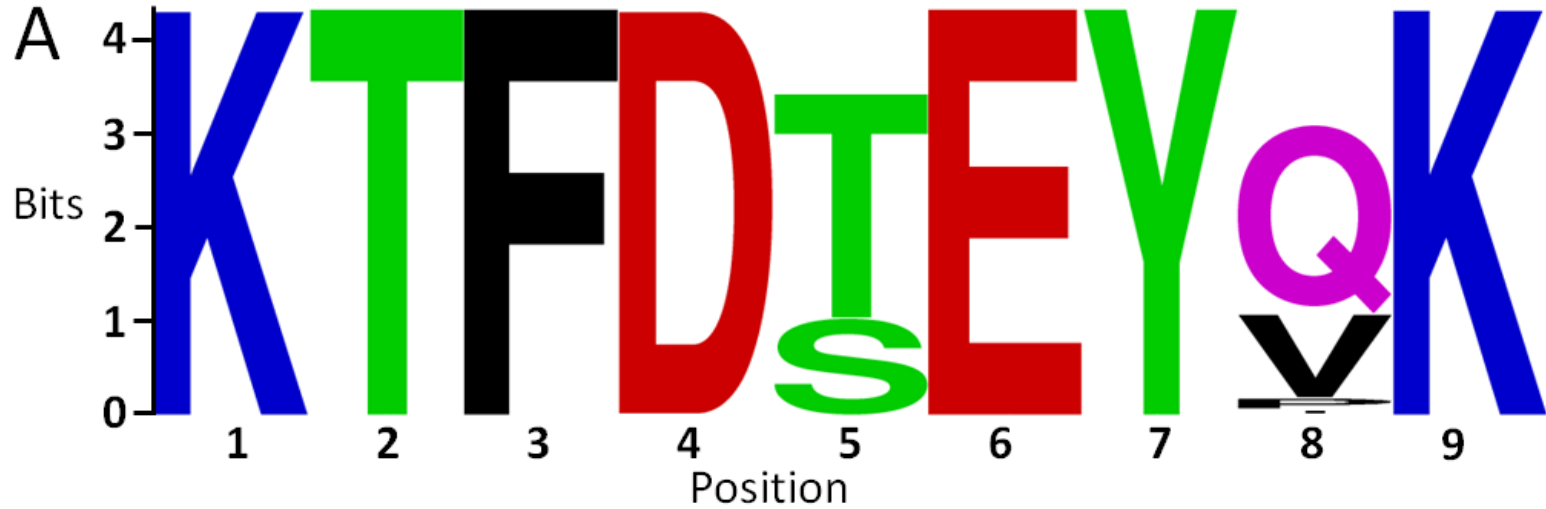
10-fold increase in the number of conserved vaccine target candidates

Block details

2722 sequences analyzed, block start: 388, word length: 9

| # | peptide | Cons(%) | Accum(%) | bit height | HLA | NetMHC predicted binding affinity |
|----|-----------|---------|----------|------------|-------|-----------------------------------|
| 1 | KTFDTEYQK | 65.03 | 65.03 | 0,84 | A0301 | 5,41 |
| 2 | KTFDSEYVK | 28.69 | 93.72 | 0,37 | A0301 | 7,19 |
| 3 | KTFDTEYPK | 3.82 | 97.54 | 0,05 | A0301 | 3,35 |
| 4 | KTFDSEYIK | 1.18 | 98.71 | 0,02 | A0301 | 8,59 |
| 5 | KTFDTEYTK | 0.29 | 99.01 | 0,00 | A0301 | 4,65 |
| 6 | KTFDSEYAK | 0.29 | 99.30 | 0,00 | - | - |
| 7 | KTFDTEYIK | 0.26 | 99.56 | 0,00 | - | - |
| 8 | RTFDTEYQK | 0.11 | 99.67 | 0,00 | - | - |
| 9 | KTFETEYQK | 0.11 | 99.78 | 0,00 | - | - |
| 10 | KTFDAEYVK | 0.07 | 99.85 | 0,00 | - | - |
| 11 | KTFNTEYQK | 0.07 | 99.93 | 0,00 | - | - |
| 12 | KTFDTEYQR | 0.04 | 99.96 | 0,00 | - | - |
| 13 | KTFDFEYIK | 0.04 | 100.00 | 0,00 | - | - |

Sequence Logo vs Block Logo



Search Flavivirus Antigen Database

Search Flavivirus records by keywords.

Choose search type

Input keyword

Refine your search for Flavivirus records

Species
Dengue virus type 1
Dengue virus type 2
Dengue virus type 3

Pathology
DF

Protein
anC
PreM
E

Strain type

Entry type

Host
Human
All Avian
All Mosquito

Search Flavivirus records using Keyword "Clone"

| Accession | Species | Entry type |
|-----------|-------------------------------|-------------------|
| FV000015 | Japanese encephalitis virus | complete proteome |
| FV000697 | Tick-borne encephalitis virus | partial proteome |
| FV000706 | Tick-borne encephalitis virus | partial proteome |
| FV001348 | Tick-borne encephalitis virus | partial proteome |
| FV001437 | Japanese encephalitis virus | partial proteome |
| FV001438 | Japanese encephalitis virus | partial proteome |
| FV001439 | Japanese encephalitis virus | partial proteome |
| FV001453 | Japanese encephalitis virus | partial proteome |
| FV001466 | Japanese encephalitis virus | complete proteome |
| FV001823 | Kunjin virus | complete proteome |
| FV001824 | Kunjin virus | complete proteome |
| FV002117 | Tick-borne encephalitis virus | complete proteome |
| FV002290 | Japanese encephalitis virus | partial proteome |
| FV002291 | Japanese encephalitis virus | partial proteome |
| FV002292 | Japanese encephalitis virus | partial proteome |
| FV002293 | Japanese encephalitis virus | partial proteome |
| FV002294 | Japanese encephalitis virus | partial proteome |
| FV002295 | Japanese encephalitis virus | partial proteome |
| FV002296 | Japanese encephalitis virus | partial proteome |
| FV002297 | Japanese encephalitis virus | partial proteome |
| FV002299 | Japanese encephalitis virus | partial proteome |
| FV002772 | Japanese encephalitis virus | partial proteome |
| FV002778 | Japanese encephalitis virus | partial proteome |

Flavivirus record FV003366

| | |
|---------------------|-------------------------------|
| Protein accession | FV003366 |
| Genebank accession | AY277665 |
| Genbank GI | 33468370 |
| Reference | 15041181 |
| Species | Dengue virus |
| Type | 1 |
| Host | Human |
| Country | Paraguay |
| Year | 2000 |
| Isolate | ARG0028 |
| Nomenclature | DENV1 Human PY ARG0028 2000 |
| Strain Type | wild type strain |
| Entry type | complete proteome |
| Proteome annotation | show/hide proteome annotation |

show/hide T cell epitopes

Predict HLA bindersProtein: Allele: Peptide length:

show/hide T cell epitopes

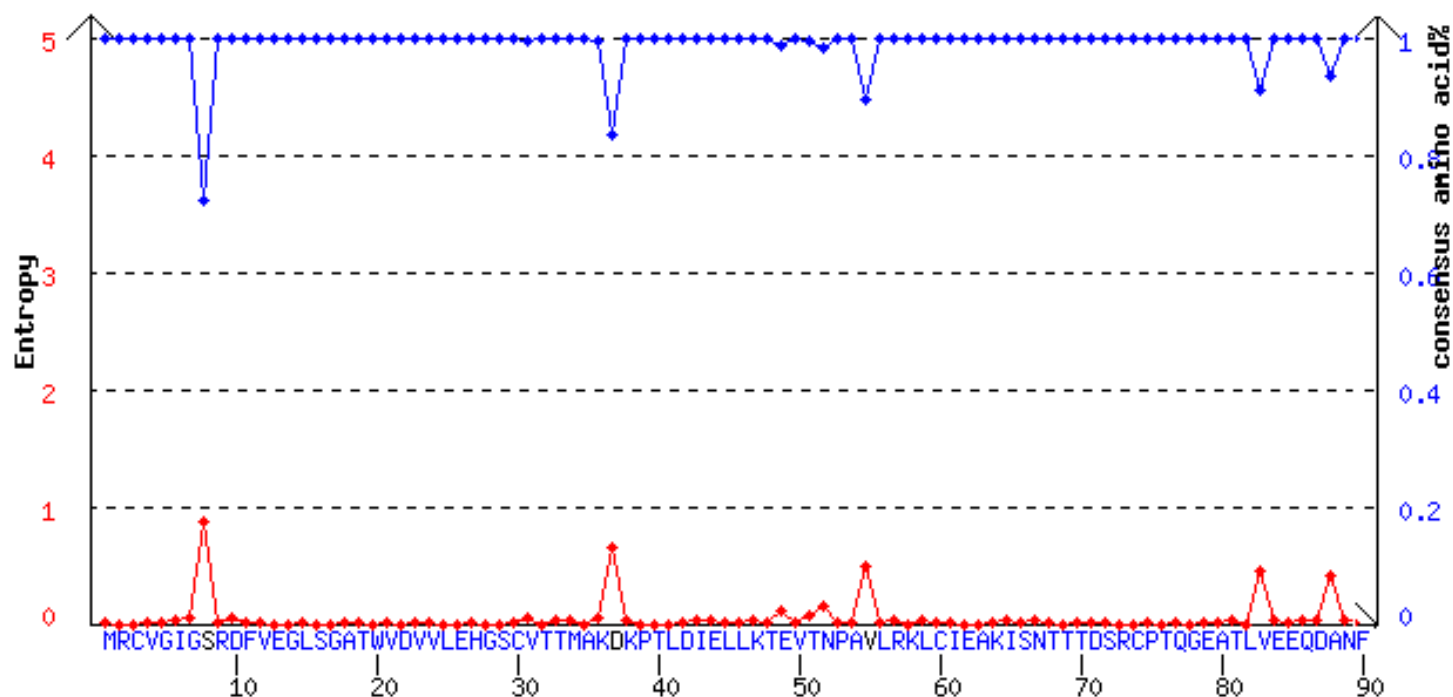
| T-cell Epitope sequence | HLA allele | Reference | in Protein |
|-------------------------|------------|-----------|--------------------------|
| ILGDTAWDFG | B7 | 15827181 | envelope protein |
| SVKLDLISY | B62 | 10453018 | nonstructural protein 3 |
| TPEGIIPAL | B35 | 7544398 | nonstructural protein 3 |
| VLMLVAHYA | A2 | 16493038 | nonstructural protein 4B |
| ILLMRTTWA | A2 | 16493038 | nonstructural protein 4B |
| MLLALIAVL | A2 | 16493038 | nonstructural protein 4A |
| FLDLPLPWT | A2 | 16493038 | envelope protein |
| GTSGSPIVNR | A11 | 12808447 | nonstructural protein 3 |
| KPWDVIPMV | B*5502 | 17626101 | nonstructural protein 5 |
| GTSGSPIVNR | A*1101 | 17565177 | nonstructural protein 3 |
| TPEGIIPAL | B35 | 18762226 | nonstructural protein 3 |
| QYSRRWCF | A24 | 16517753 | nonstructural protein 3 |
| EIVLMCHAT | DPw2 | 7690424 | nonstructural protein 3 |
| EIVLMCHAT | DPw2 | 9568963 | nonstructural protein 3 |
| VLMCHATFT | DPw2 | 9568963 | nonstructural protein 3 |
| NREGKIVGLYGNGW | DRB1*1501 | 16081844 | nonstructural protein 3 |
| PTRVVAEMAEALKG | DRB1*1501 | 16081844 | nonstructural protein 3 |

Sequence variability analysis result

Consensus sequence is shown below X axis with conserved positions in blue. A conserved position must fulfill: entropy ≤ 1 , gap fraction ≤ 0.1 , and consensus amino acid% $> 90\%$.

Entropy and % of consensus amino acid at each position based on 1574 Human Dengue virus type 1 complete envelope sequences in the database

[Download full result file](#) | [Download consensus sequence file](#)



[show/hide multiple sequence alignment of Human Dengue virus type 1 envelope proteins](#)

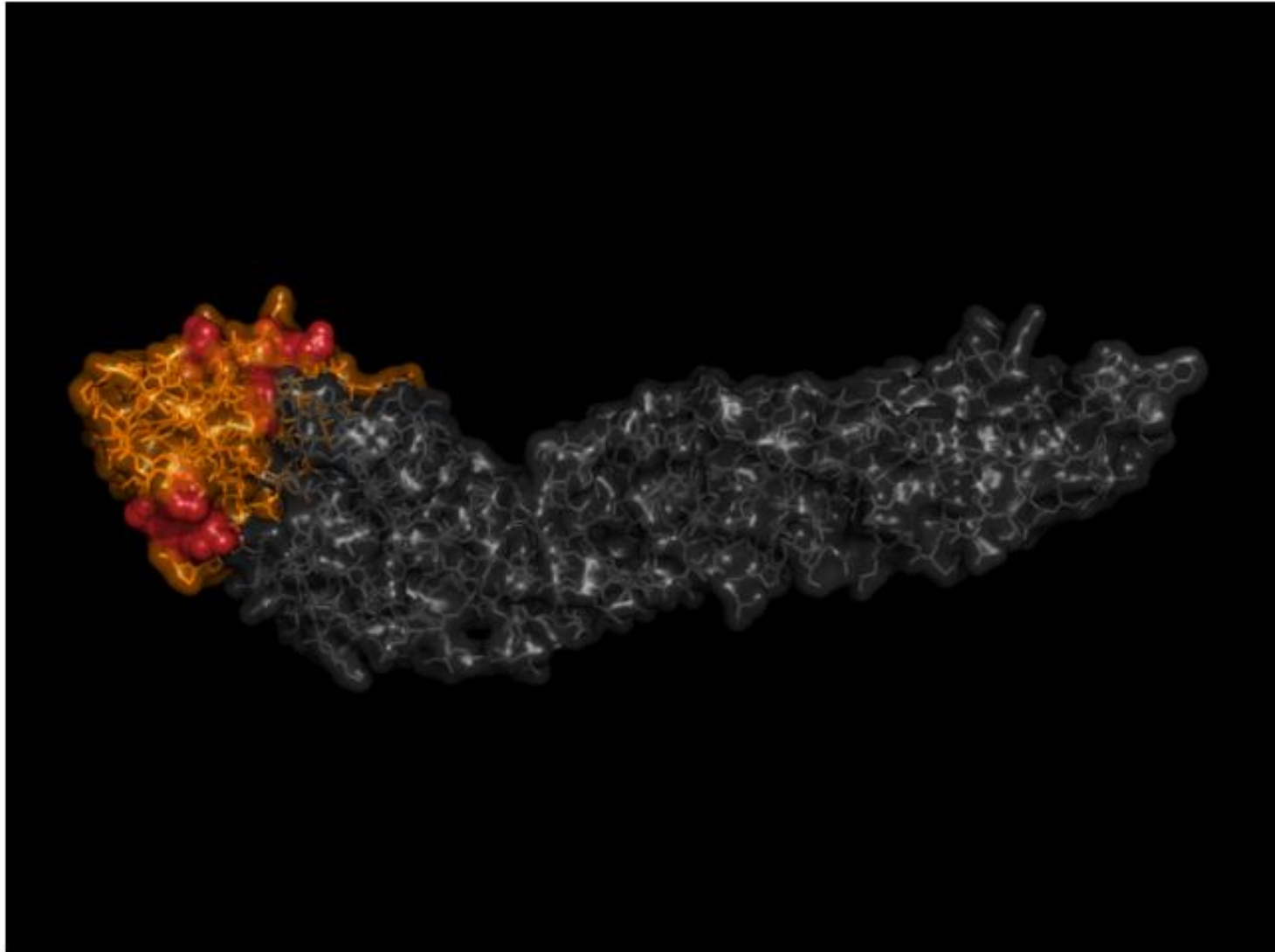
B-cell epitope summary table for 4 dengue virus subtypes (Amino acids highlighted at the positions where neutralizing epitopes appear more often than non-neutralizing epitopes)

| pos | -- | W101 | G104 | G106 | L107 | F108 | K174 | P176 | W231 | Y299 | V300 | T301 | L301 | N302 | T303 | F304 | G304 | K305 | V305 | L306 | F306 | K307 | L308 | K308 | V308 | E309 | V309 |
|-----|------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| DV1 | neutralizing | | | | | | | | | 1 | 1 | | | | | | | | | | | 1 | 1 | | | 2 | |
| DV1 | non-neutralizing | 1 | | | 1 | 1 | | | | | | | | | | | 3 | 1 | | | | | | | | | |
| DV1 | Unknown | | | | | | | | | | | | | | 1 | | 4 | 2 | | | | 2 | | | | | |
| DV2 | neutralizing | 9 | 9 | 10 | 3 | 1 | | | 1 | 1 | 1 | | | | 1 | | 7 | 8 | | | 5 | 11 | | | 5 | 1 | 5 |
| DV2 | non-neutralizing | | | | | | | | | | | | | | 1 | | | | 1 | 1 | | 1 | | 1 | | 1 | |
| DV2 | Unknown | 1 | | 5 | 5 | 1 | | | | | | | | | | | 2 | | | | | 1 | | | | | |
| DV3 | neutralizing | | | | | | | | | | | 4 | 1 | 8 | 3 | 8 | | | 7 | 8 | | 2 | | 16 | | 8 | |
| DV3 | non-neutralizing | 1 | | | 1 | 1 | | | | | | | | | | | 3 | 2 | | | | 1 | | | | | |
| DV3 | Unknown | | | | | | | | | | | | | | 1 | | 3 | 1 | | | | 1 | | | | | |
| DV4 | neutralizing | 1 | 1 | 1 | | | 2 | 2 | | 1 | 1 | | | | | | | | | | | | | | | | |
| DV4 | non-neutralizing | 1 | | | 1 | 1 | | | | | | | | | 6 | | 4 | 2 | 6 | 6 | | 7 | | 6 | | 7 | |
| DV4 | Unknown | | | | | | | | | | | | | | 1 | | 3 | 1 | | | | 1 | | | | | |

Input key neutralizing positions (seperated by comma) of your interest. They will be highlighted on 3D structure.

3D structure

Key positions are in red.



SUMMARY

Introduction

Definitions

Framework

Bioinformatics tools

Examples

Malaria

FLAVIdB

Conclusions

CONCLUSIONS

Traditional databases are being supplemented by knowledge-based systems

KBS integrate data, knowledge base, analytical tools, and workflows for multiple automated analyses

KBS suggest experiments and learn from results

Acknowledgements

Guanglang Zhang PhD – system design and implementation

Lars R. Olsen MSc – FLAVIdB

