

GlaxoSmithKline

From genome to vaccine

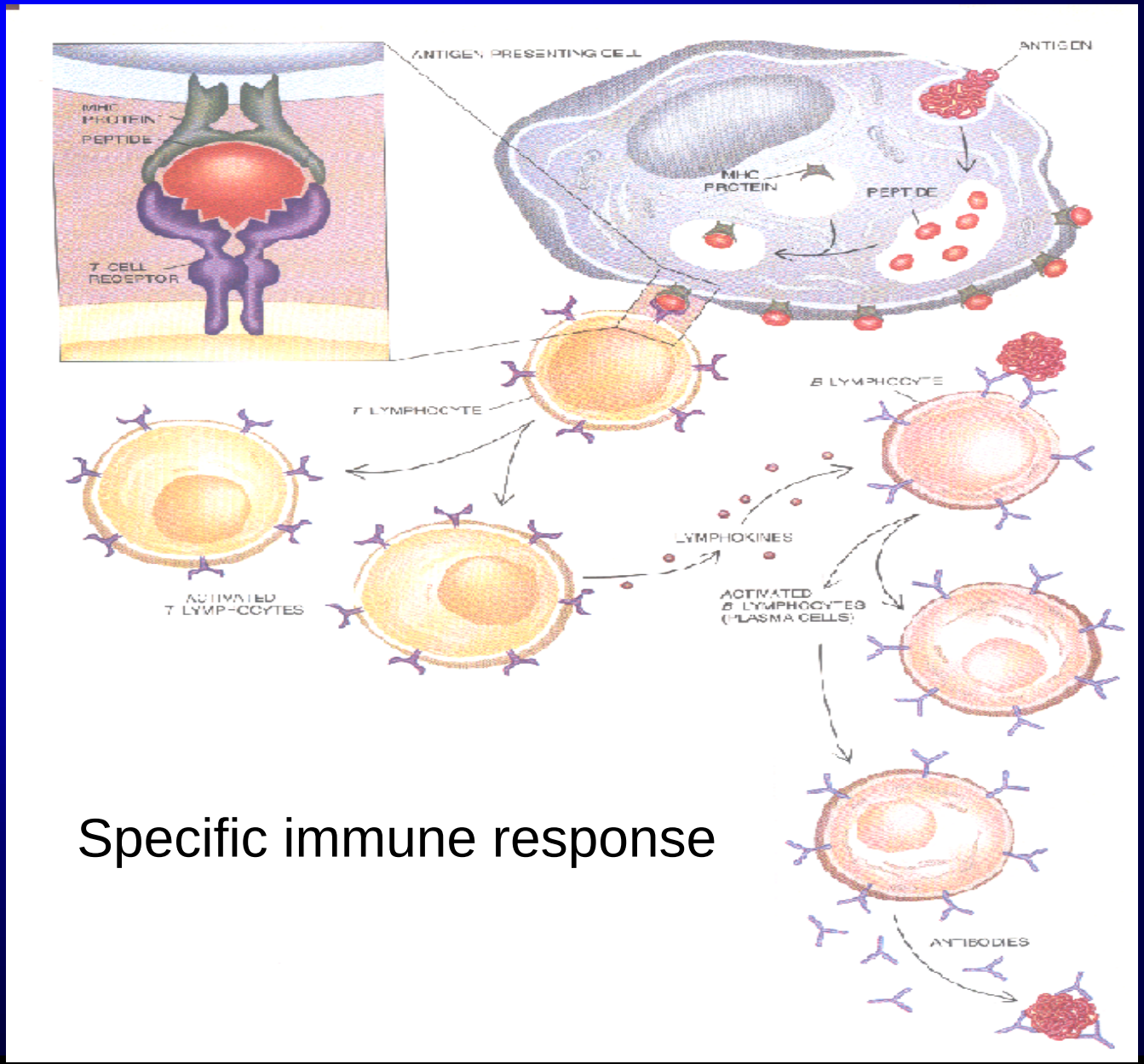
"in silico" bacterial genome mining for new

anti-infective vaccine development

Joelle Thonnard 11 March 2002 Master in bioinformatic KUL

- Vaccines
- In silico genome mining
  - bacterial genomes
  - target proteins
  - mining tools
- Bioinformatic protein analysis
- Conclusion

gsk GlaxoSmithKline



Specific immune response

GlaxoSmithKline

Vaccine concept

- Administration of altered preparation of pathogen
- to induce a specific immune response involving B-cells (antibody production) and T-cells (helper and cytotoxic cells)
- enabling next encounter with the full pathogen to induce a secondary response, more rapid and effective than the primary response

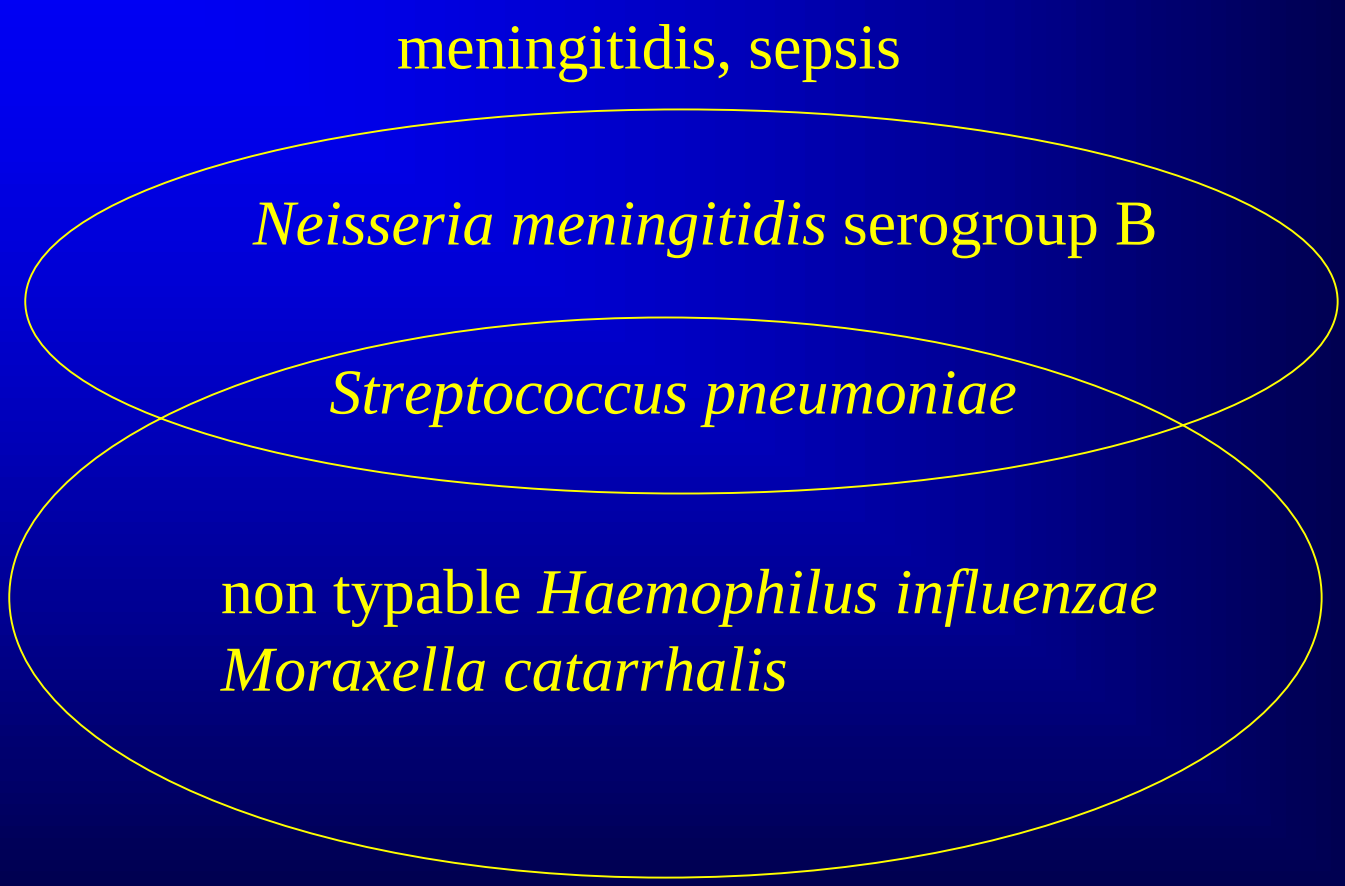
gsk GlaxoSmithKline

Current bacterial vaccines

- **Toxins**: diphteric toxin, tetanic toxin
  - only possible for some vaccines
- **Killed whole cell**: B.pertussis
  - inflammatory reaction
- **Purified proteins**: acellular B. pertussis
- **Capsular polysaccharides (PS)**: Hib, N. meningitidis A and -C, S. pneumoniae
  - not always possible or sufficient

gsk GlaxoSmithKline

Need for protein-based vaccines



GlaxoSmithKline

Neisseria meningitidis

- Capsulated gram negative bacteria
- Meningitidis, sepsis
- > 15 serogroups
- 3 serogroups clinically important

A (Africa)

C (EU/USA)

B (EU/USA)

Capsular polysaccharide (PS) based vaccines available

Blebs-based vaccine available strain specific (Cuba)

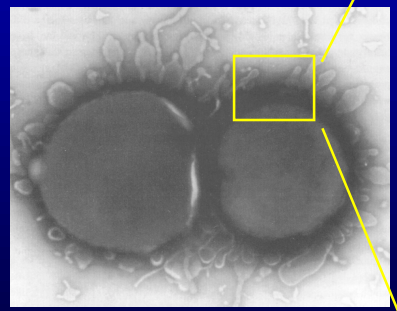
gsk GlaxoSmithKline

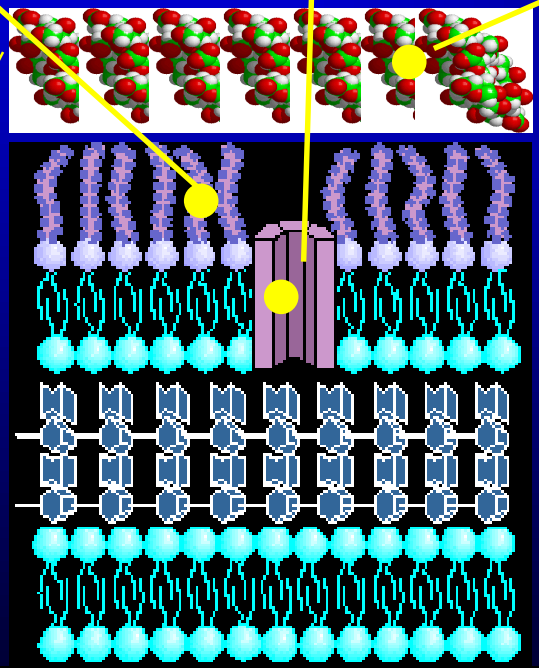
N. meningitidis type B

LPS
toxic, variable

OMP
major Ag = variable
minor Ag = conserved

Capsular PS
autoimmunity





Capsular polysaccharides — **Capsule**

LPS/LOS
Phospholipids — **Outer-membrane**

Peptidoglycan — **Periplasm**

Phospholipids — **Cytoplasmic membrane**

Bacterial cell wall

gsk GlaxoSmithKline

M. catarrhalis, non typable *H. influenzae*

- Gram negative bacteria
- No capsule
  - no capsular polysaccharides to use as vaccine compounds
- No vaccine available

gsk GlaxoSmithKline

Streptococcus pneumoniae

- Gram positive bacteria
- Capsular PS : > 80 serotypes
- Current vaccines
  - 23-valent capsular polysaccharides
    - protective in adults for blood stream infection but not pneumoniae
    - infants and young children poor responder to PS
  - 7-valent capsular polysaccharides conjugated to a carrier protein to enhance T-cell reponse
    - 7 PS included cover 80% of bacteriemia and 65% of otitis media

gsk GlaxoSmithKline

Candidate vaccine proteins

- Surface exposed proteins
  - Target bacteria mainly extra-cellular
  - Antibodies directed against surface exposed components shown to be protective
    - anti-PS antibodies are protective for Hib MenA-C
  - T-cell reponse probably involved in protection but few data available (for extra-cellular bacteria)

gsk GlaxoSmithKline

New vaccine development

- New antigen discovery
  - Classical laboratory work
  - Highthrouput experimental system
  - Genomic mining
  - Microarrays
  - Proteomics



- Antigen production
  - Construction proposal
  - Expression of recombinant proteins



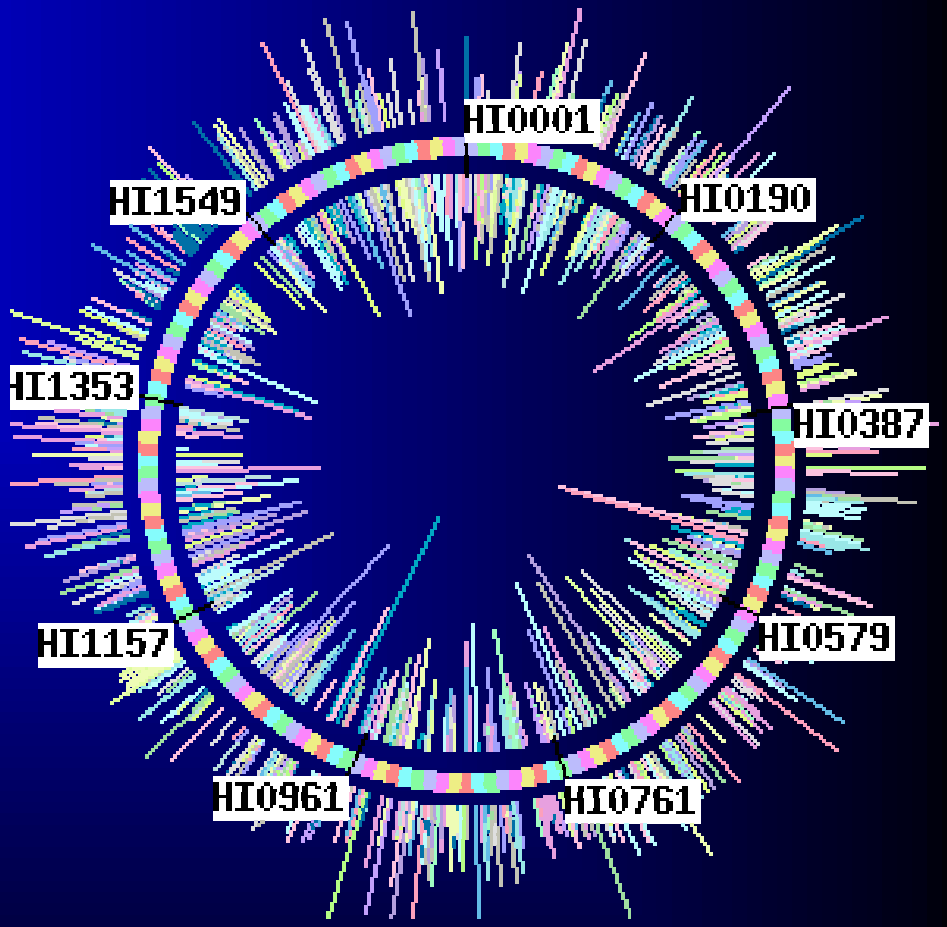
- Test protective potential in animal models

gsk GlaxoSmithKline

Bacterial genomic effort start

1995

- first complete genome sequenced
- *Haemophilus influenzae* Rd strain
- 1,830,138 bp
- 1709 proteins



GlaxoSmithKline

Bacterial genomic effort today

March 2002

- eubacterial genomes
  (human pathogens & other)
  - > 57 fully sequenced
  - ~ 115 in progress or nearly finished
- several strains of the same species
  - *Staphylococcus aureus* : 7 strains sequenced
- additionnal private genomic projects

gsk GlaxoSmithKline

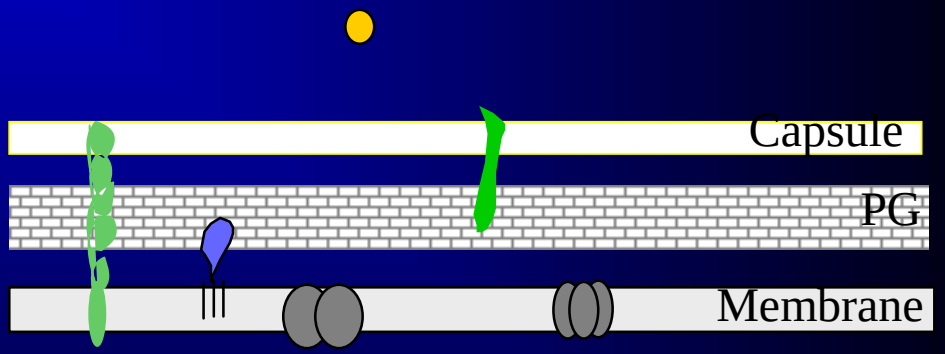
Genomes mined

- *N*eisseria *meningitidis* type B
  - 2,2 Mb 2158 ORFs
- *Moraxella catarrhalis*
  - 1.9 Mb ~1500 ORFs
- nt *Heamophilus influenzae*
  - 1.6 Mb ~1800 ORFs
- *Streptococcus pneumoniae*
  - 2.1 Mb 2236 ORFs

gsk GlaxoSmithKline

Surface exposed proteins in Gram +

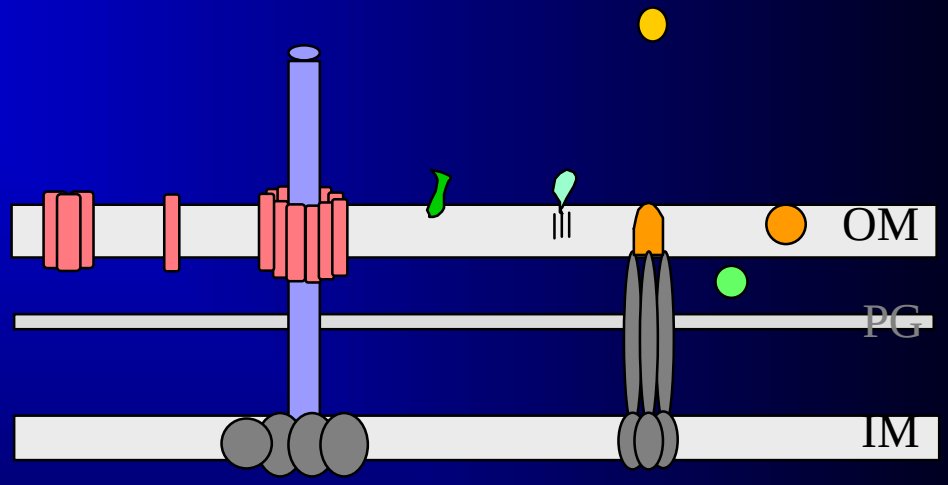
- Proteins anchored to peptidoglycan
  - LPxTG motif, …
- Proteins anchored in the membrane
  - transmembrane α-helix
- Lipoproteins
  - type 2 signal sequence
- Secreted proteins
  toxins
  proteases
  - type 1 signal sequence ...



gsk GlaxoSmithKline

Surface exposed proteins in Gram -

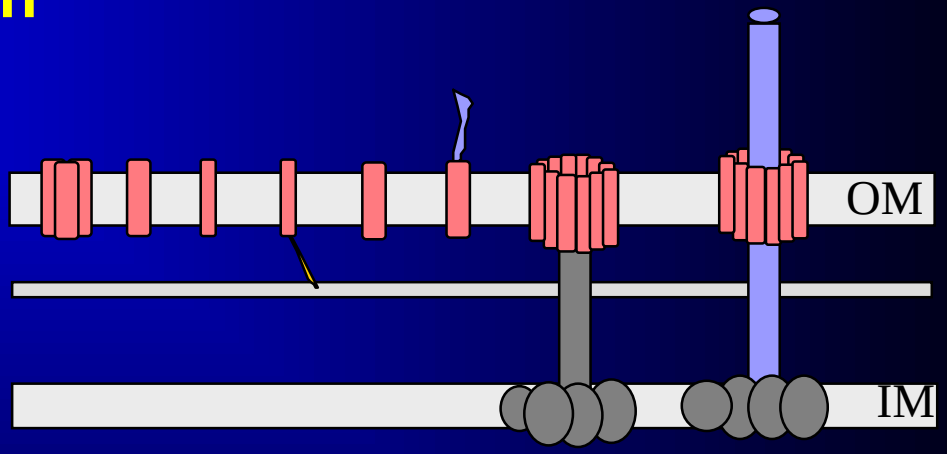
- Outer membrane proteins
  - **porin or porin-like intregral OMP**
  - **pilins**
  - **not piliated adhesins**
  - **OM lipoproteins**
  - **others**
- Secreted proteins
  - **toxins**
  - **proteases**
- Periplasmic proteins
  - **some potential immuno-protective proteins**



gsk GlaxoSmithKline

Porin or Porin-like iOMP functions

- porins for ions, glucose
- porins involved in iron acquisition
- adhesins
- structural proteins
- enzymes
- autotransporters
- secretion systems
- pili apparatus



gsk GlaxoSmithKline

Search strategies

- ORFs predited or retrieved from the web
- ORF automatic annotation
- Identify candidates by homology
  - Identification of proteins in the genome that are homologous to known protective proteins in other bacteria
- Identify candidates by features
  - Identification of all potential surface exposed proteins
    - identification of specific features of surface exposed proteins
    - combination of methods to identify & select ORFs

gsk GlaxoSmithKline

ORF prediction

- Genemark
- Glimmer

- problems with unfinished genomes

gsk GlaxoSmithKline

Automatic annotation

- In house Perl script
  - launch of a series of analyses
  - summary of results
    - html pages (automatically)
    - acces database (semi-automatically)
  - no automatic interpretation of the results

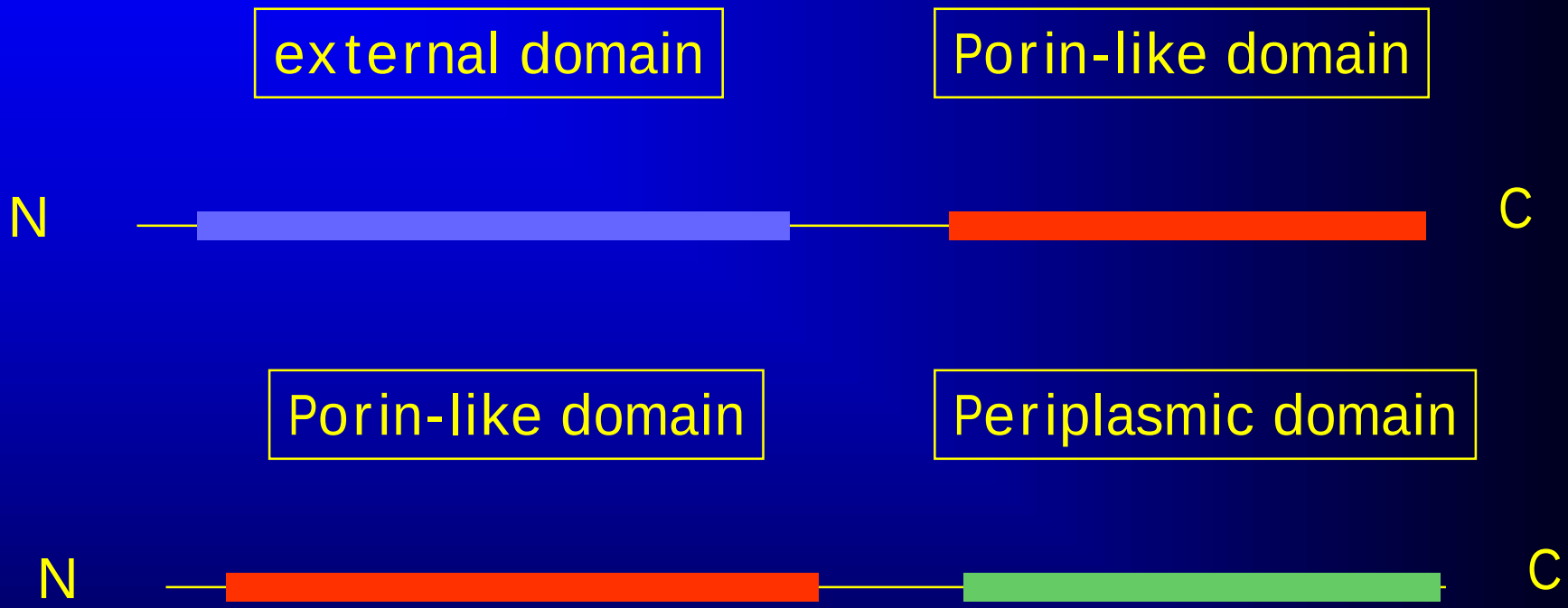
gsk GlaxoSmithKline

Search for known homologs

- comparison of the ORFs of the genome to known proteins from genebank, swissprot, trembl ...
  - blast
  - 1st hits
    - often unknown protein
    - protein annotated by homology
    - multi-domain protein
  - > interpretation problem

gsk GlaxoSmithKline

Frequently multi-domain proteins



gsk GlaxoSmithKline

Search for conserved domains

- Prosite
  - biologically significant sites, patterns and profiles
  - findpattenrs (GCG)
- Pfam
  - Hidden Markov models based on multiple alignments of protein domains
  - hmmsearch
- Blocks
  - motifs
  - blocksplus
- Prints
  - a fingerprint is a group of conserved motifs used to characterise a protein family
  - blocksplus

gsk GlaxoSmithKline

Type 1 signal prediction

N- (charged) H- (hydrophobic) C- domains

MKTLKTLLAVSASSLLAMSANA<>

- spscan (GCG)
  - von Heijne matrices
  - Mac Geoch
- signalp
  - neural network

gsk GlaxoSmithKline

Type 2 signal prediction

N- (charged) H- (hydrophobic) C- domains

MKIKALGVVLLASSMA**L**AG<>**C**

- findpatterns
  - LxxC
  - prosite pattern

    {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C

gsk GlaxoSmithKline

α-helix transmembrane regions

- regions spanning
  - membrane of gram+ bacteria
  - inner-membrane of gram- bacteria
- hydrophobic α-helix
  - tmhmm

gsk GlaxoSmithKline

2D structure prediction

- DSC
- DSC multi
- Psipred
- NPS@
  - consensus computed from predictions of several methods
  - SOPM, HNN, DPM , DSC , GORIV, PHD , PREDATOR and SIMPA
  - http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html

gsk GlaxoSmithKline

Localisation prediction

- NNSPL
  - neural network
  - FOR GLOBULAR PROTEINS
- Locpred
  - Markov chain models
  - FOR GLOBULAR PROTEINS

gsk GlaxoSmithKline

Automatic interpretation not sufficiently reliable

- Example : Genequiz / Lion
  - automatic interpretation based on search of meaning hits, keywords ...
  - failure in 1/6 cases compared to careful manual annotation
- Reasons
  - multidomain proteins
  - some homologs automatically mis-annotated

gsk GlaxoSmithKline

Identify candidates by homology

- Single sequences as query
  - known immuno-protective proteins in other bacteria
  - comparison to the mined genome
  - blast, fasta
- HMM as query
  - HMM derived from multiple alignments containing known immuno-protective proteins
  - comparison to the mined genome
  - Prodom
  - Hmmer
  - more sensitive

gsk GlaxoSmithKline

Identify candidates by features

- Example 1
  porins Gram -
- Example 2
  peptidoglycan associated proteins Gram +
- Example 3
  lipoproteins

gsk GlaxoSmithKline

Porin gram -

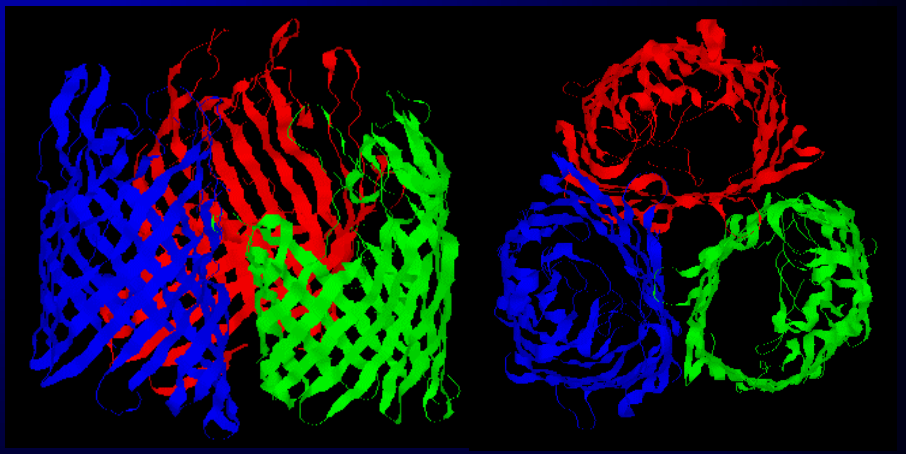
- Type 1 N-terminal signal sequence (usually)

  N- (charged) H- (hydrophobic) C- domains

  MKTLKTLLAVSASSLLAMSANA<>

- aromatic C-terminus amino acid (usually)

- porin domain
  β-barrel
  β-strands
  8 to 26 (even number)
  amphipatic
  shirmer picks (aromatic AA)



gsk GlaxoSmithKline

Porin gram - prediction

- type 1 signal sequence
- mainly β-barrels predicted in a window of 100 AA
  - window needed as multi-domain proteins
- amphipatic β-strands
- shirmer picks
- C-term aromatic amino acid
- localisation prediction

⇒ Prediction based on combination of criteria

gsk GlaxoSmithKline

Peptidoglycan associated proteins gram +

LPxTG H- (hydrophobic) B- (basic)

-KEPLPDTGSEDEANTSLIWGLLASLGSLLLFRRKKENKDKK

- Variable length extracellular domain N-term
- conserved hexapeptide
- cluster of basic residues

gsk GlaxoSmithKline

Peptidoglycan associated proteins gram + prediction

- LPxTG regular expression
  - not specific
- prosite regular expression L-P-x-T-G-[STGAVDE]
  - more restrictive
- HMM derived from multiple alignment of Prodom
  - less restrictive
  - more specific

gsk GlaxoSmithKline

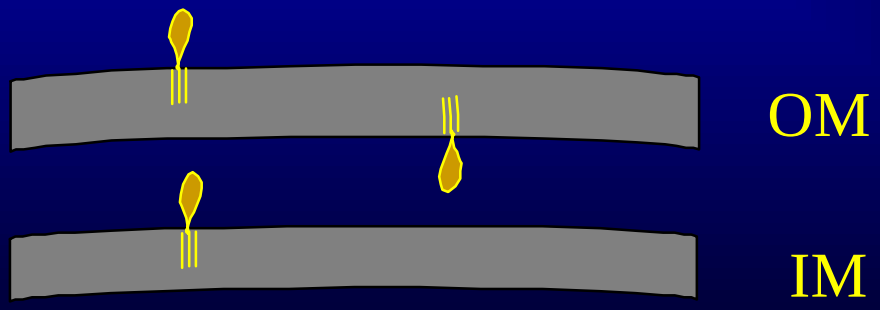
Lipoprotein features

- Type 2 N-terminal signal sequence

N- (charged) H- (hydrophobic) C- domains

MKIKALGVVLLASSMA**L**AG<>**C**

- no reliable feature associated with inner or outer membrane location



gsk GlaxoSmithKline

Example of successful results

- *N. meningitidis* serogroup B genome (Chiron)
  - 350 proteins selected for expression
  - test in animal models
  - 7 candidates identified

gsk GlaxoSmithKline

Example of successful results

- *M. catarrhalis* genome (GSK)
  - expression and testing of 35 proteins
  - 5 candidates still under evaluation

gsk GlaxoSmithKline

Protein analyses

- Variability
- B-cell epitopes
- T-cell epitopes
- Domains

gsk GlaxoSmithKline

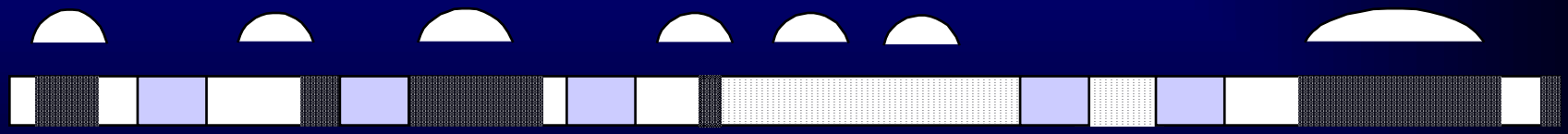
Variability study

- Important to select as vaccine compound proteins conserved and present in all virulent strains
  - Compare selected genes in the different genomes when several genomes of the same species available
  - Compare selected genes or proteins to sequences of sequence databases

gsk GlaxoSmithKline

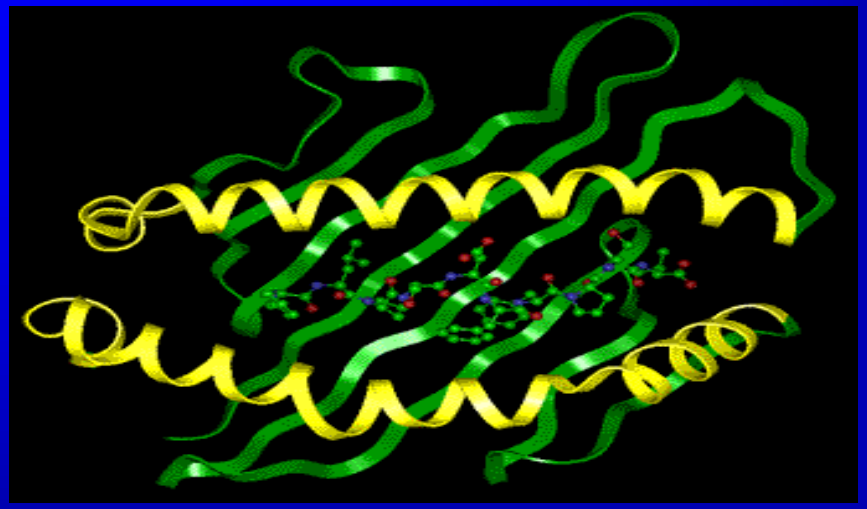
B-cell epitope prediction

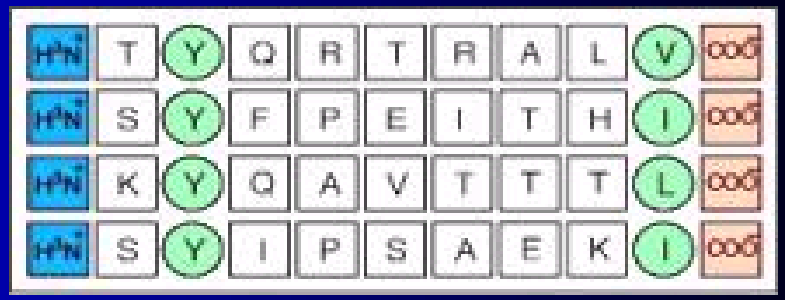
- Exposed mobile regions of protein
- 3D-structure based prediction
- multiple sequence alignment
- 2D-structure prediction
- hydrophobicity / hydrophilicity profile
- Plotstructure Jameson-Wolf method
- PHD neural network



gsk GlaxoSmithKline

T-cell epitope prediction

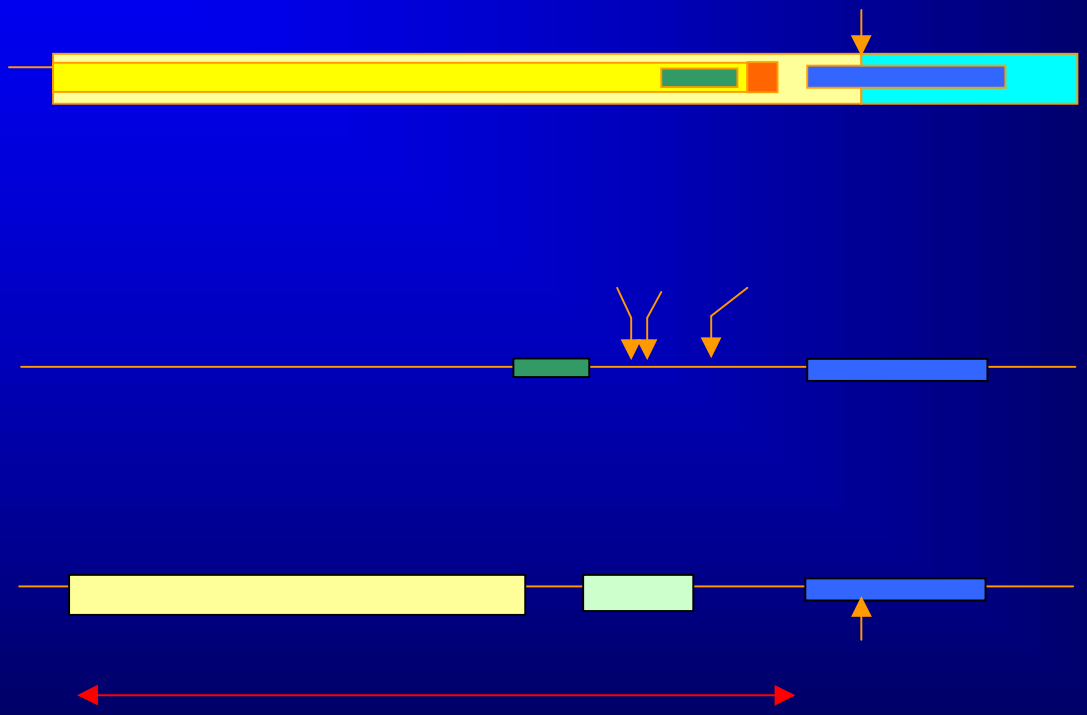




- Several HLA-binding prediction methods
- not yet fully reliable depending on alleles however reduction of peptides to test

gsk GlaxoSmithKline

Domain prediction



GlaxoSmithKline

Conclusion

- Pros & Cons
  - when genomic data available, rapid tool to identify new potential antigens / epitopes
  - don’t allow to identify new class of protective antigens
  - validation step still limiting
  - complementary tool to experimental methods

gsk GlaxoSmithKline