From genetic coding

to

protein structure prediction

May 2014

Jean-Luc Jestin

Département de Biologie Structurale et Chimie
Institut Pasteur

Biological chemistry
Molecular biology

Mathematics

Quantitative approaches

Arithmetic
Basic number theory

Genetic code

Error-correcting codes
Coding theory

A method to identify rules between protein sequences and structures

OUTLINE

1. *Basic number theory,*
   the p-adic distance
   and applications

2. *Symmetries in the genetic code*

3. *Applications in protein structure prediction*

Conclusion

Representation of numbers by quadratic forms

X^2 represents 1, 4, 9 and 25 over Q

but not 2 and 3

X^2 and Y^2 represent the same numbers over Q

but not X^2 and $2Y^2$

Hasse-Minkowski theorem

For two quadratic forms to represent 0 over Q,

it is necessary and sufficient that

they represent 0 over Q_p for all p and over R

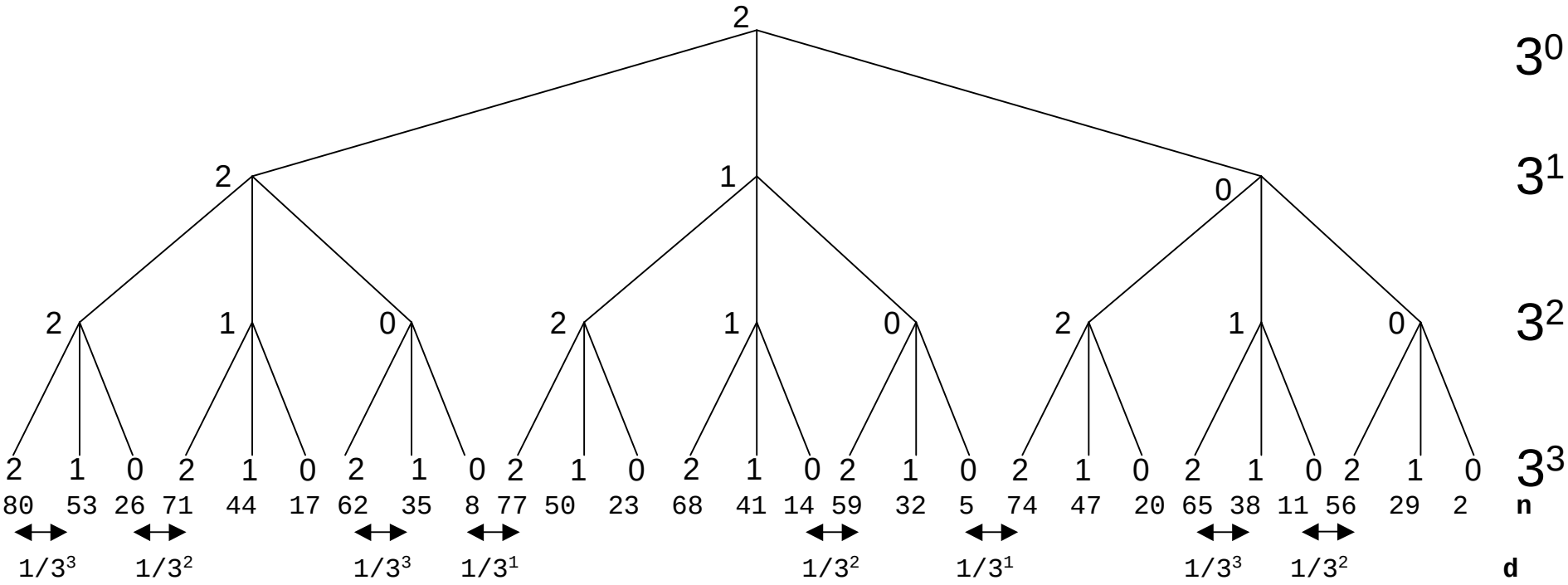
Metrics

$$\text{norm}(a+b) \leq \text{norm}(a) + \text{norm}(b)$$

For ultrametric distances such as the p-adic distance, there is a stronger inequality:

$$\text{norm}(a+b) \leq \sup(\text{norm}(a); \text{norm}(b))$$

The 3-adic distance and a 3-adic tree

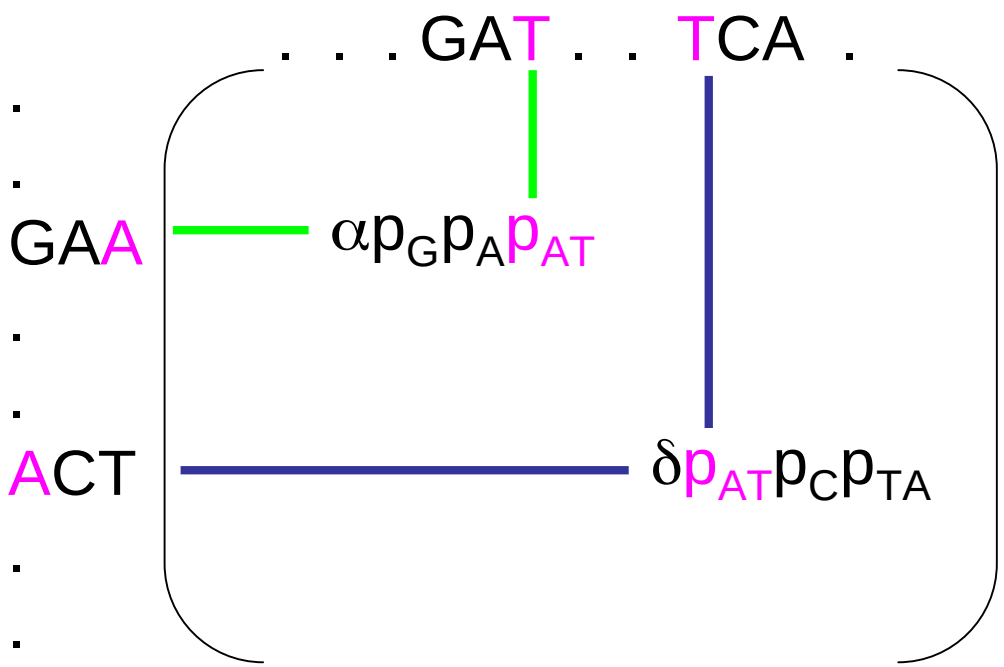


Phylogeny:

The smaller the distance, the higher the number of ancestors in common.

In: « DNA replication and mechanism » Nova, 2011

The codon substitutions matrix M



The « frequentist interpretation » of probabilities in statistical physics

Dill, Pressé et al., Rev. Mod. Phys. 2013

The same mutation, a factor in common, p-adically close numbers:
the p-adic distance

Symmetries by base substitutions of degeneracy in the genetic code
are symmetries of the classes of quadratic forms over Q_p

Biosystems 2010

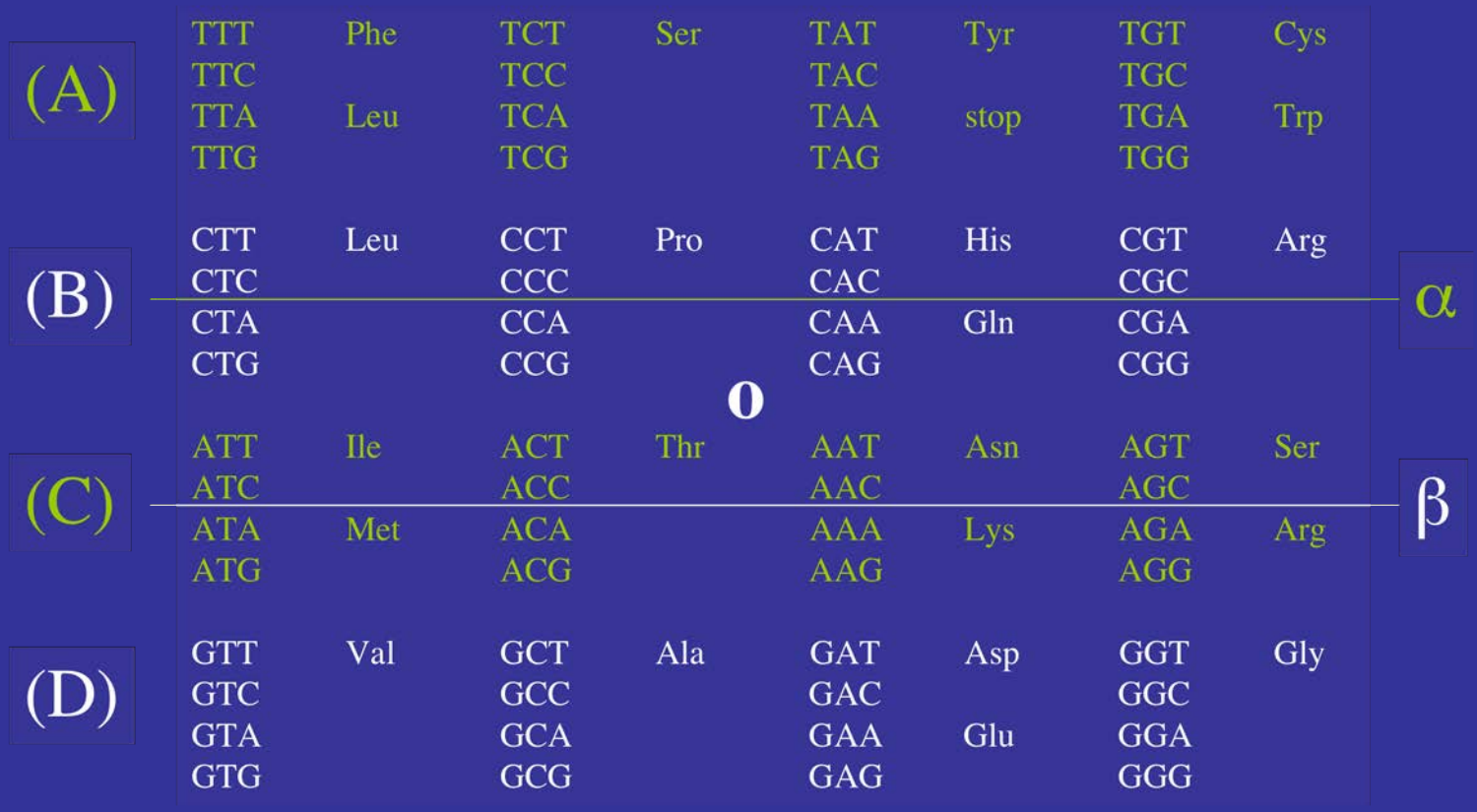
1. *Basic number theory, the p-adic distance and applications*

2. *Symmetries in the genetic code*

A representation of the genetic code

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC		TCC		TAC		TGC	
TTA	Leu	TCA		TAA	stop	TGA	stop
TTG		TCG		TAG		TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Gln	CGA	
CTG		CCG		CAG		CGG	
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	Lys	AGA	Arg
ATG	Met	ACG		AAG		AGG	
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	Glu	GGA	
GTG		GCG		GAG		GGG	

Symmetries of degeneracy by base subtitutions



Comp. Rend. Biol. 2006
Rumer, Proc. Natl. Acad. USSR 1966

Genetic code codon assignments
preserve information
minimize the deleterious effects of mutations

Mutations	**Conserved property**	**Reference**
Transitions at the 3d base	*Amino acid*	*Goldberg & Wittes Science 1965*
Transversions at the 3d base	*Metabolic pathway*	*Wong PNAS 1975*
Transversions at the 1st base	*Degeneracy*	*C. R. Biol. 2006*
Single-base substitution leaving T as the 2d base	*Hydrophobicity*	*Woese PNAS 1965*
Single-base deletion within stop codons	*Genes: full-length ORF*	*with Kempf FEBS Letters 1997*
Single-base substitution for side chains with square mass	*Kinetic energy over* Q *(not* R*)*	*with Guilloux Biosystems 2012*
Three-bases substitutions (GT/AC, AG/CT, GT/AC)	*2' or 3' –OH tRNA-aminoacylation*	*with Soulé J Theor Biol 2007*

Kinetic energy conservation by mutation

protein evolution by mutation

Amino acid *1* with side-chain mass m

$$E_1 = \frac{mX^2}{2} \qquad (1)$$

Amino acid *2* with side-chain mass M

$$E_2 = \frac{MY^2}{2} \qquad (2)$$

if $M = mA^2$ then $E_2 = \frac{m(AY)^2}{2}$

and same numbers represented by *(1)* and *(2)* over Q,
i.e. kinetic energy disturbances are minimized during evolution consistently with the codon arrangement

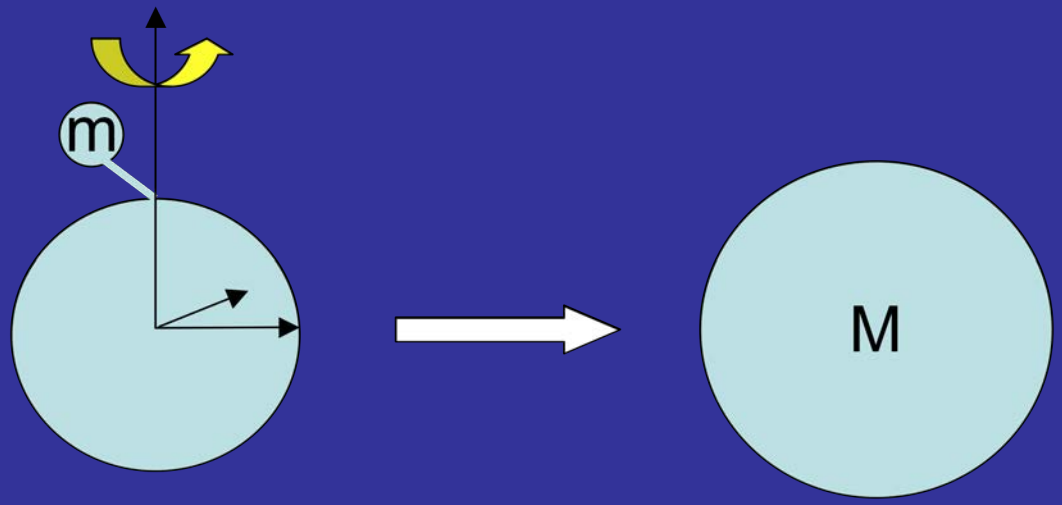
with A.Guilloux, Biosystems 2012

1. *Basic number theory, the p-adic distance and applications*

2. *Symmetries in the genetic code*

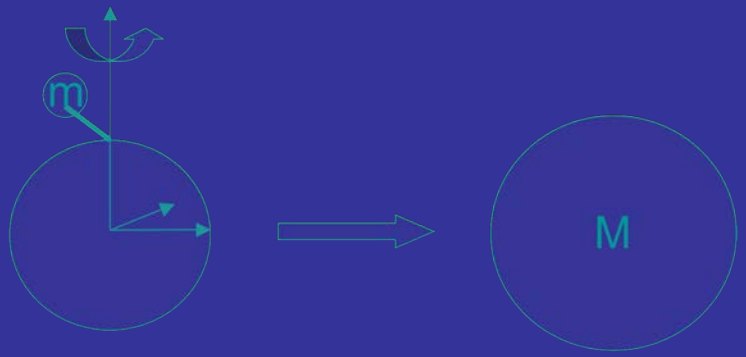
3. *Applications*
   in protein structure prediction
   for beta-sheets

Elementary protein folding step



with A.Guilloux & B.Caudron, Comp. Struct. Biotech. J. 2013

Total energy conservation law during the elementary folding step

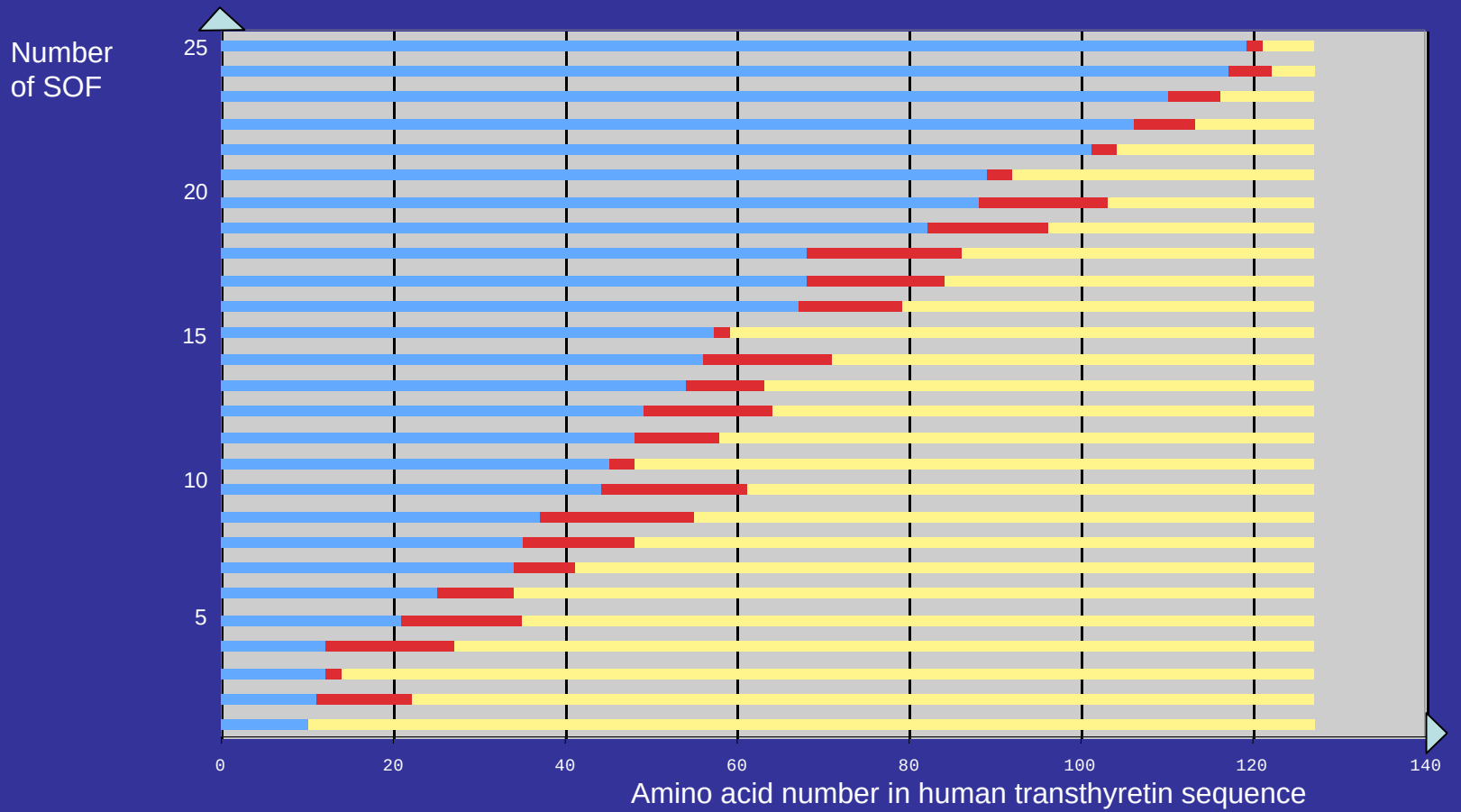


$$\frac{mX^2}{2} = \frac{MY^2}{2} + E$$

There is an infinite number of solutions (X,Y) over Q, $\forall E$
if m/M is a square

with A.Guilloux & B.Caudron, Comp. Struct. Biotech. J. 2013

Sequences optimized for folding



SOF length *(red)* ≤ 20
Protein length *(x-axis)* ≤ 250

using pdb2 and pdb23 on mobyle.pasteur.fr

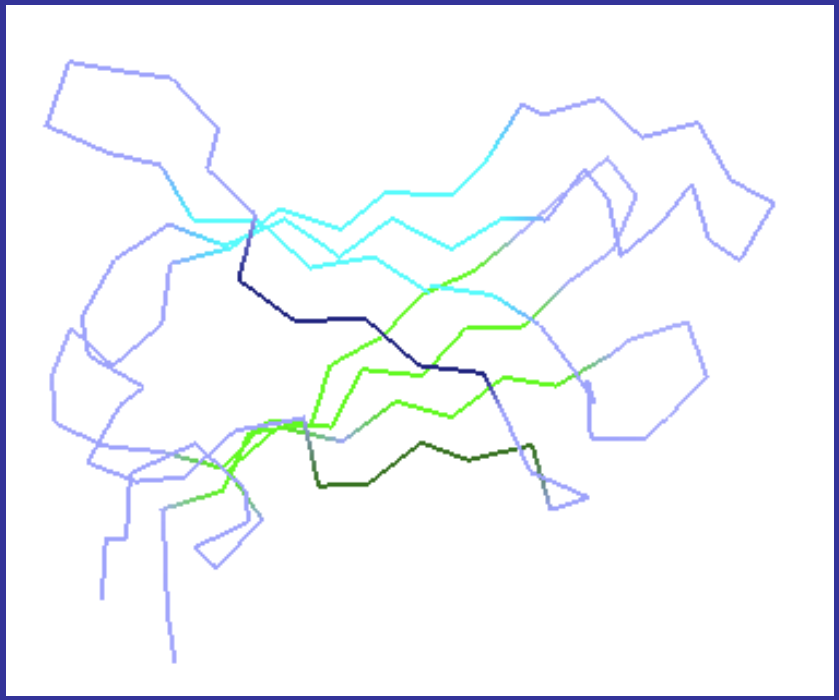
with A.Guilloux & B.Caudron, Comp. Struct. Biotech. J. 2013

Edge strand prediction

Open beta-sheets
with 4 strands or more
more than 3 amino acids-long

Two (training / test) sets

Rule: edge strands predicted
from the sequence
by extreme mean number of SOFs
(Sequences Optimized for Folding)
except if 2D-loops



Human transthyretin structure

96 predictions
p-value $< 9 \times 10^{-7}$
prediction accuracy 74-78%
improved using multiple sequence alignments

with A.Guilloux & B.Caudron, Comp. Struct. Biotech. J. 2013

Edge strand prediction methods

Structural bioinformatics & physical biochemistry methods:

strand length, hydrophobicity

Sternberg & Thornton J. Mol. Biol. 1977
Von Heijne & Blomberg J. Mol. Biol. 1978

protein folding, total energy conservation

edge strand prediction accuracy: 74-78%

with A.Guilloux & B.Caudron, Comp. Struct. Biotech. J. 2013

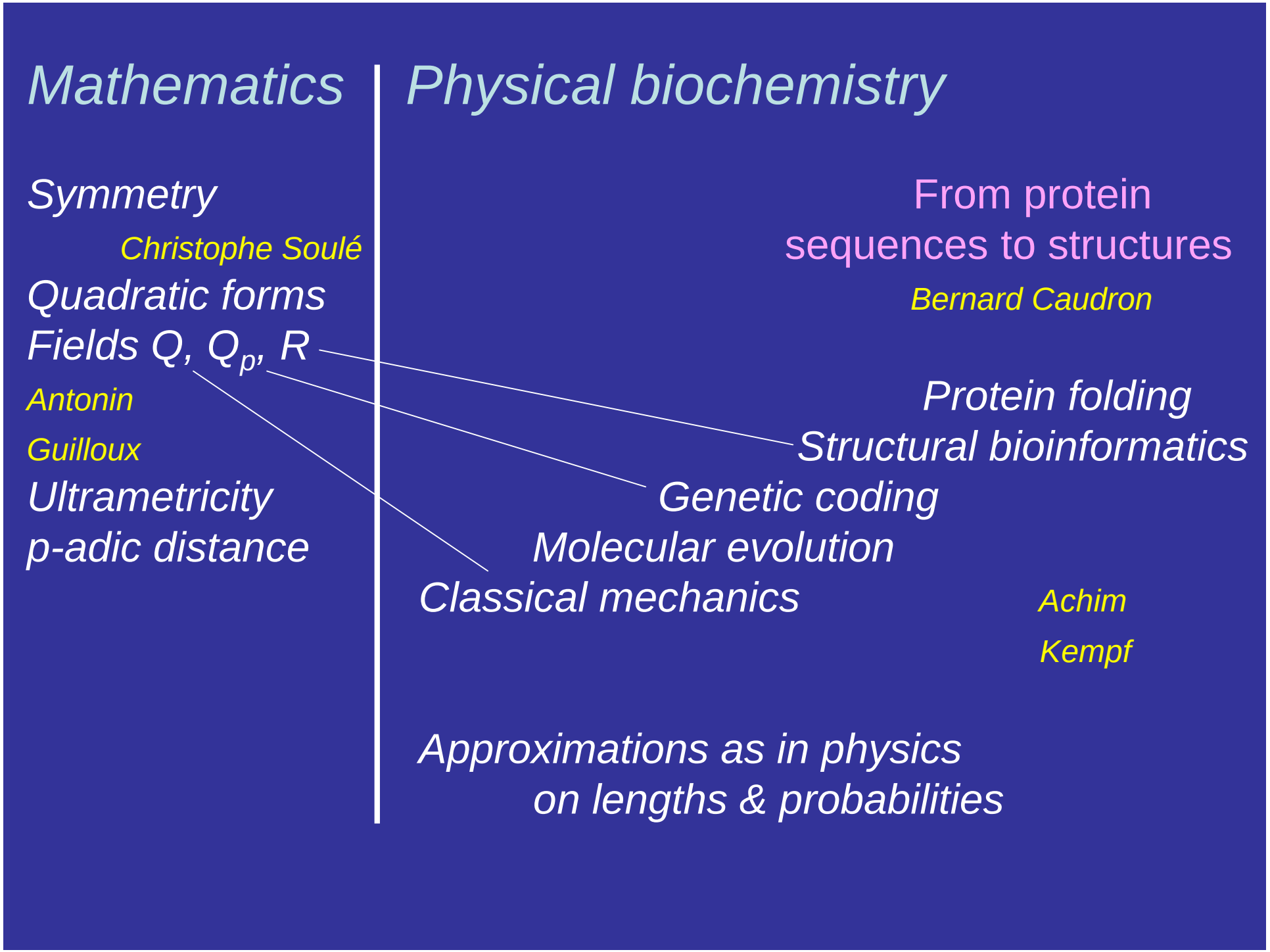
Machine learning methods:

support vector machines

decision tree algorithms

edge strand prediction accuracy: 70-83%

Westhead et al. Prot. Sci. 2003
Faulon et al. J. Mol. Model. 2006


Mathematics
Physical biochemistry
Symmetry
Christophe Soulé
Quadratic forms
Fields Q, Q_p, R
Antonin
Guilloux
Ultrametricity
p-adic distance
From protein
sequences to structures
Bernard Caudron
Protein folding
Structural bioinformatics
Genetic coding
Molecular evolution
Classical mechanics
Achim
Kempf
Approximations as in physics
on lengths & probabilities

Acknowledgements

Achim Kempf
Department of Applied Mathematics
Waterloo University, Canada

Christophe Soulé
Institut des Hautes Etudes Scientifiques, Bures
Centre National de la Recherche Scientifique, France

Antonin Guilloux
Institut de Mathématiques de Jussieu, Paris
Ecole Normale Supérieure de Lyon

Bernard Caudron
Centre d'Informatique pour la Biologie
Institut Pasteur, Paris