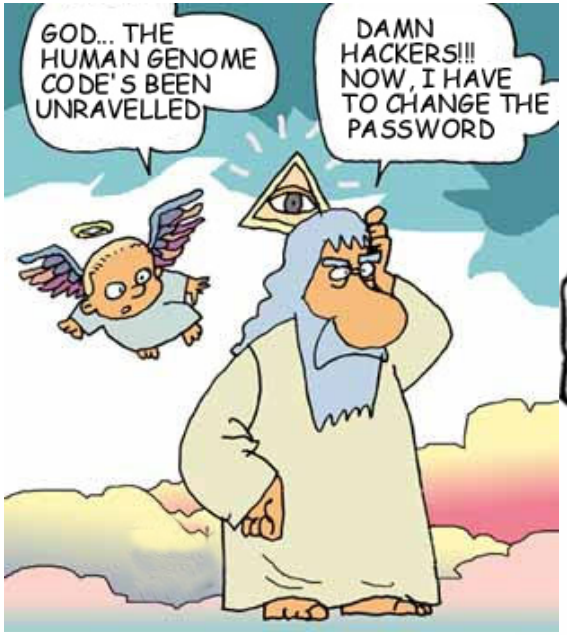
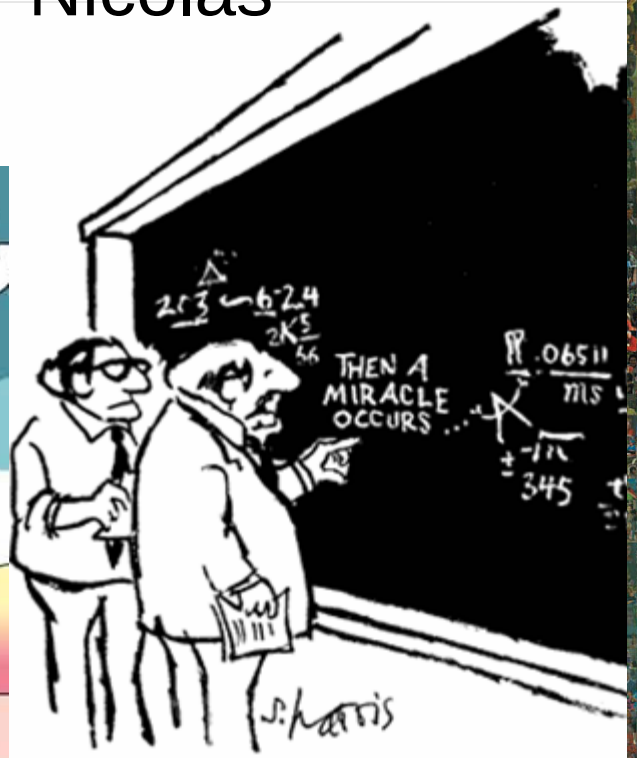
Syntaxe, Raisonnement et Génomes

Jacques Nicolas





"I think you should be more explicit here in step two."

Soutenance HDR

13 Mai 2008

IRISA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

centre de recherche RENNES - BRETAGNE ATLANTIQUE

Symbiose

La biologie en quelques mots …

- Systèmes vivants = systèmes symboliques :
  code, messages, stockage d'information, discrétisation, interactions typées.
  Niveau informationnel $\leftrightarrow$ Niveau Biochimique

- Systèmes vivants= systèmes complexes :
  nombreux types d'objets, nombreuses interactions, la diffusion est réduite aux petites molécules.
  Multiplication des échelles d'abstraction

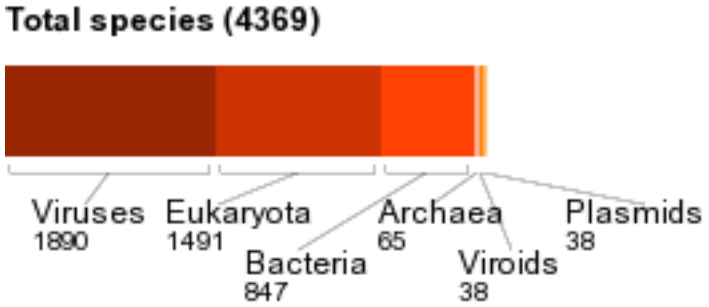
- Essor technologique et accumulation des données très au-delà des progressions théoriques.
  Source de modèles originaux

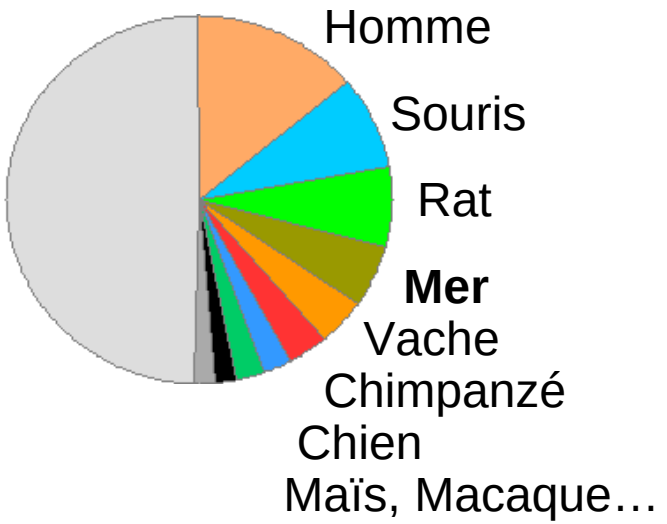
La biologie, c'est beaucoup de mots

- 242 GB de séquences brutes EMBL avril 2008; Génomes complets : 77 eukaryotes, 630 bactéries, 54 archées.

- Passage du séquençage d'espèces représentatives au séquençage de populations (espèces proches, métagénome, individus…)

- La molécule d'ADN est formée de deux brins complémentaires à orientation inversée





Mon travail : Jeux de mots Langages formels et combinatoire

- Analyse lexicale : Observer les séquences avec le bon niveau d'abstraction; *Forest, Pygram, Modulome*
- Analyse syntaxique : Permettre au biologiste de modéliser des structures sur ses séquences; *Stan*, *Logol*
- Inférence grammaticale : Obtenir automatiquement des modèles sur des familles de séquences. *Grammaires algébriques, Automates d'états finis*

+ de la Biologie !

Analyse lexicale

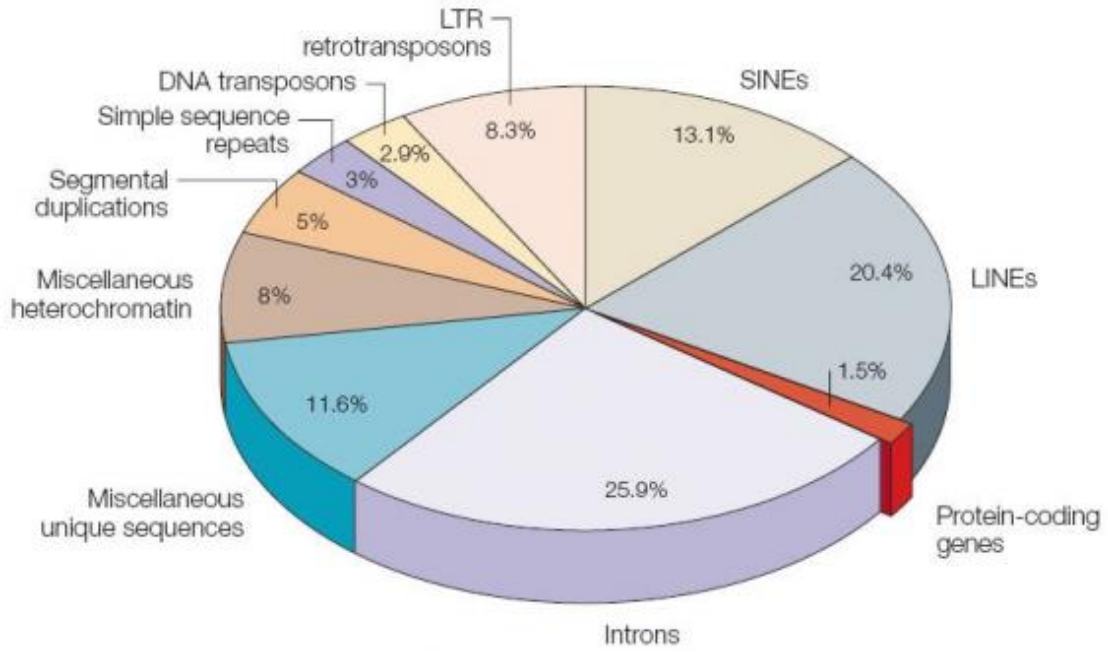
Le lexique: Evolution = Répétitions + Mutations ponctuelles + Réarrangements

- Les génomes sont des entités extrêmement dynamiques. Mécanismes essentiels de l'adaptation : duplications, mutations et réarrangements chromosomiques.

- Chez les vertébrés : 1 duplication / gène x 100M années (1/3 gènes dupliqués chez l'homme ou chez la levure. Exemple des récepteurs olfactifs >1000gènes); 0.3 substitution /base x 100M années).

- Duplications en tandem ou dispersées, à des échelles très variées (quelques bases, 1 seul gène, des gènes contigus, 1 chromosome-aneuploïdie-, 1 génome -polyploïdie-…)

Quelques exemples de répétitions connues



Composition du génome de l'homme (~3Gb)

- Transposons/retrotransposons : éléments mobiles
  - Short Interspersed Nuclear Elements (*Alu*: ~0.3 Kb, 10M copies)
  - Long Interspersed Nuclear Elements (entre 0.5 et 5Kb, 200k copies)
  - ...
- Duplications segmentales : erreurs de réplication (en tandem direct, entre 5 et 100kB)
- Satellites
  - hétérochromatine : ~40 b, 1K copies, 300Kb en tout.
  - Télomères, minisat : ~10b, 30 copies, 5Kb en tout
  - STR, microsat : 1 à 4b, 20 copies, 100b en tout
- Familles de gènes

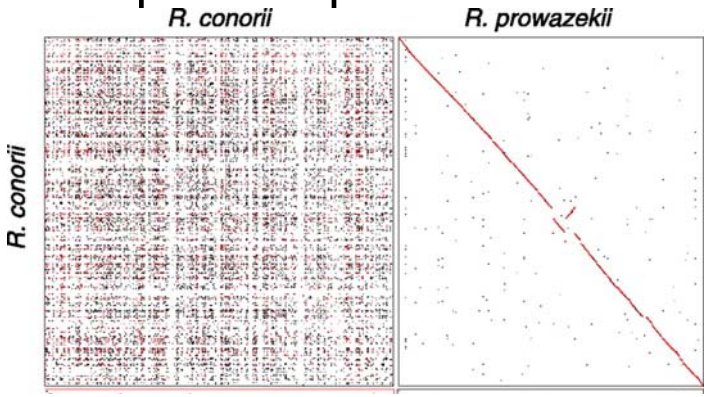
Analyse lexicale : visualiser les facteurs

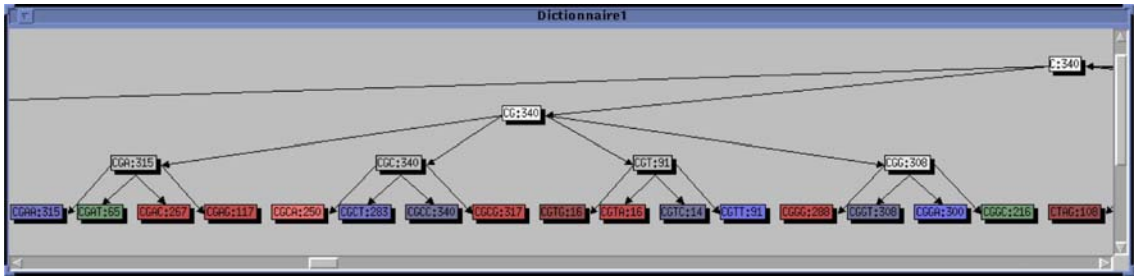
Forest

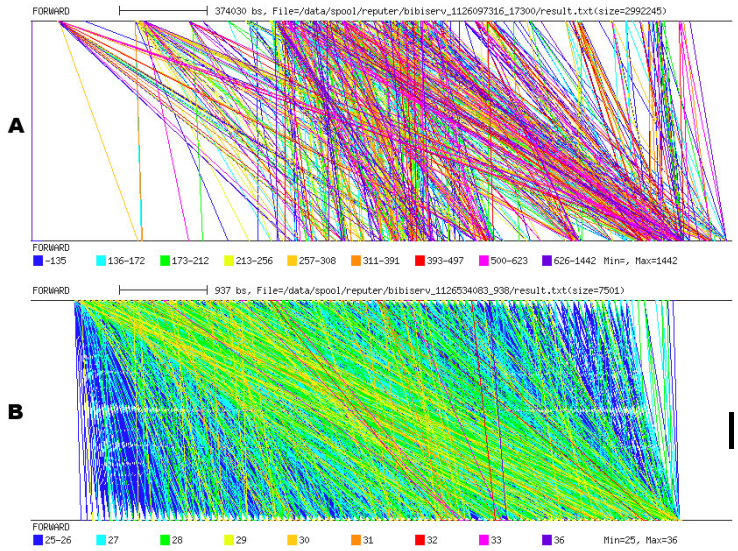
Représentations usuelles quadratiques

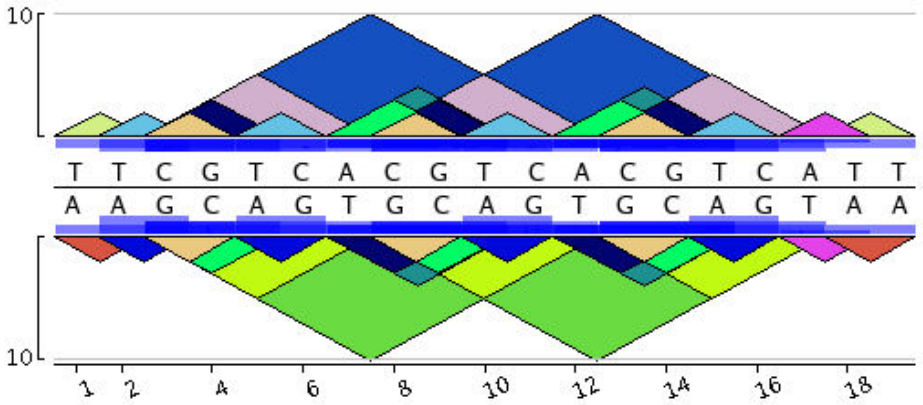
Index : Arbre des suffixes + Calcul d'attributs

1997 Thèse Robin Gras









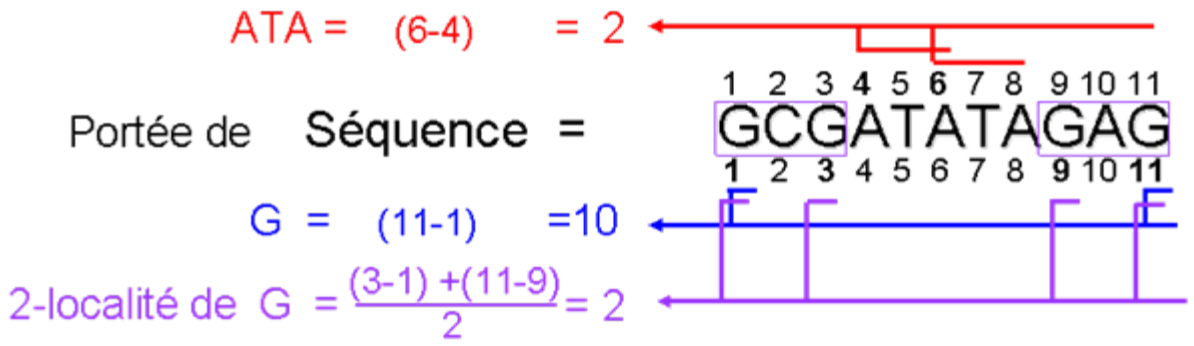
Pygram

Index+ Répétitions maximales + Hiérarchie

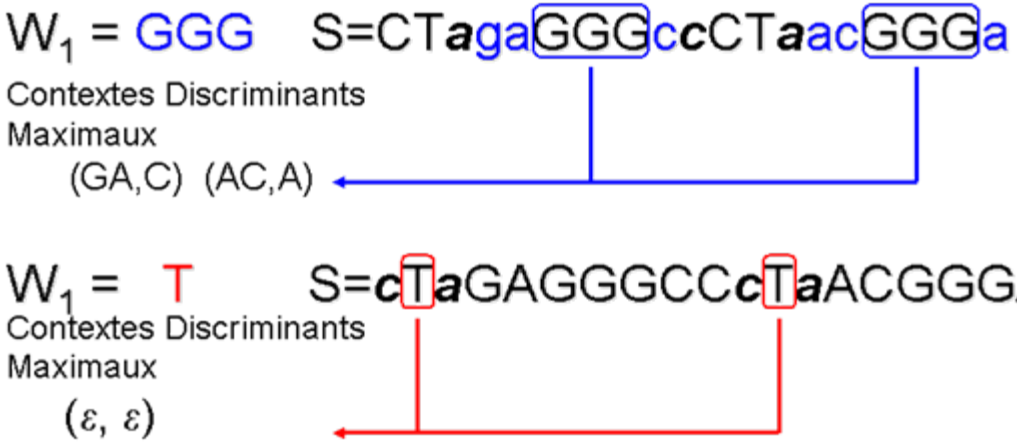
P. Durand BMC Bioinformatics 2006, **7:**477

Localité et maximalité des répétitions : deux concepts essentiels en biologie

- Localité : distribution concentrée des répétitions en clusters



- Maximalité : contextes des répétitions non similaires



Le projet ANR Modulome: découper un génome en tranches

(LM2E Ifremer Brest, LEPG CNRS Tours, Institut Jacques Monod Paris)

Décrypter le non codant à différents niveaux d'abstraction.

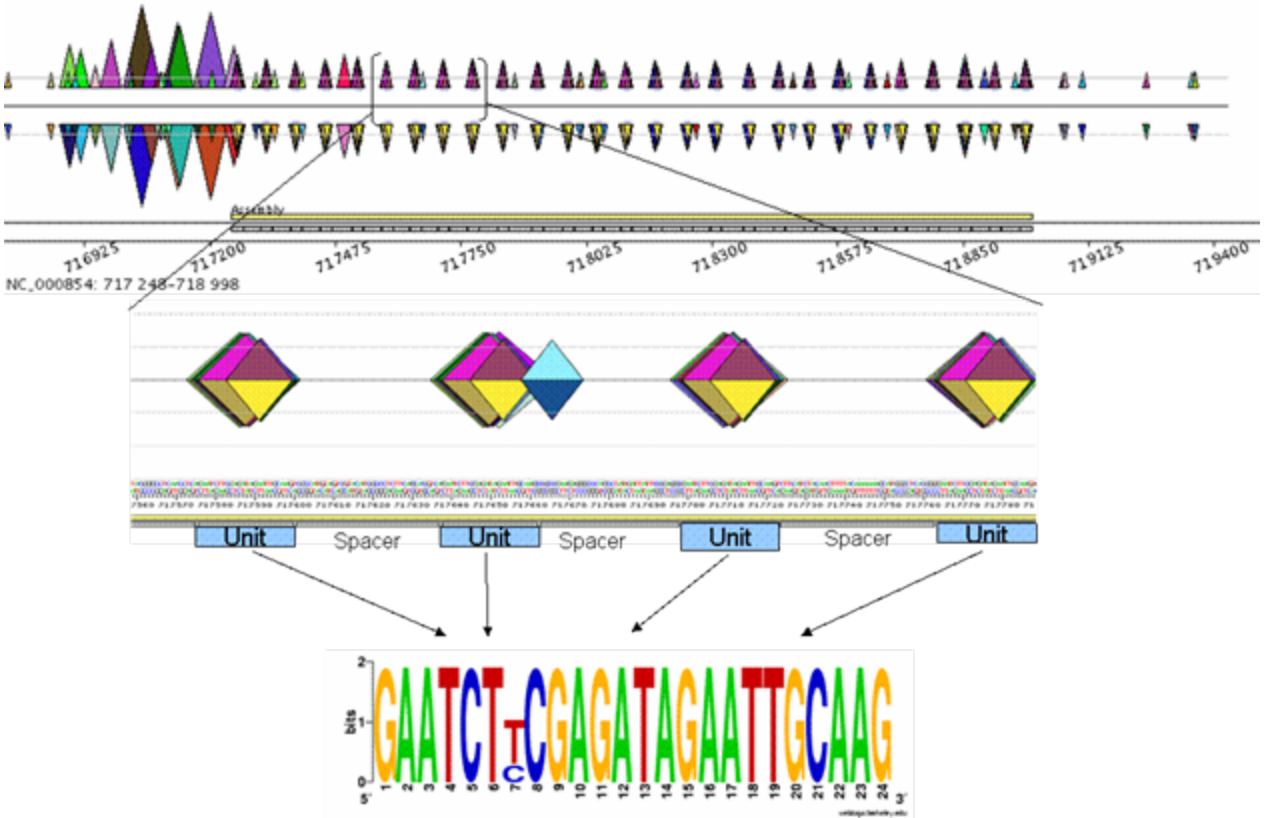
- Analyse d'éléments génétiques importés dans les archées et les bactéries (transfert horizontal), les CRISPR, une structure de défense microbienne;
- Analyse d'éléments génétiques mobiles : Hélitrons, MishMar1 et autres transposons;

C. Rousseau, A. Siegel, F. Coste, P. Durand, F. Mahé, S. Tempel

2006-2009 En soumission : Journal of algorithms + Genome Biology

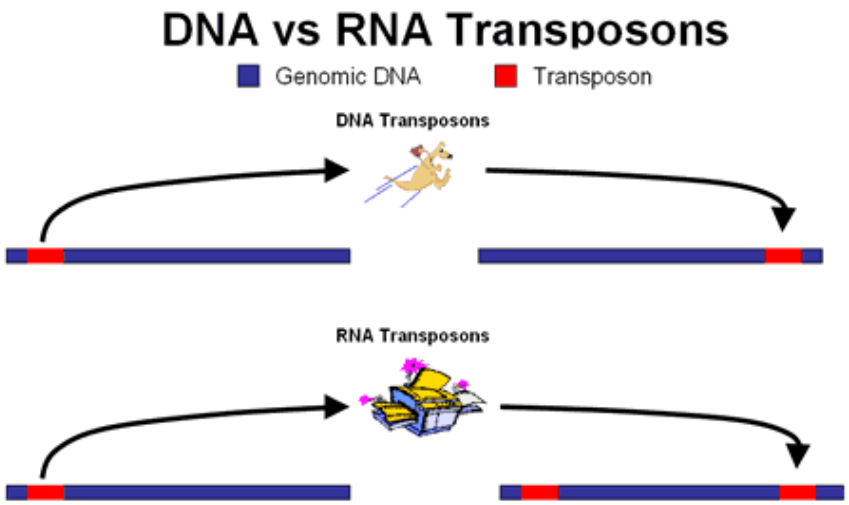
Recherche de CRISPR

- Répétitions locales et maximales pour repérer les CRISPR : beaucoup moins de paramètres que les méthodes actuelles;
- Répétitions τ–voisines pour repérer indirectement les gènes associés.



Les transposons

- En 1983, Barbara McClintock reçoit le prix Nobel de Médecine pour la découverte de transposons dans le maïs



- Génomes Poulet : 27% Homme : 45% Maïs : 60% Lys : 99%
- Rôle dans l'adaptation, la régulation, le transfert de gènes, emploi croissant en génomique fonctionnelle (cancer…)

DomainOrganizer

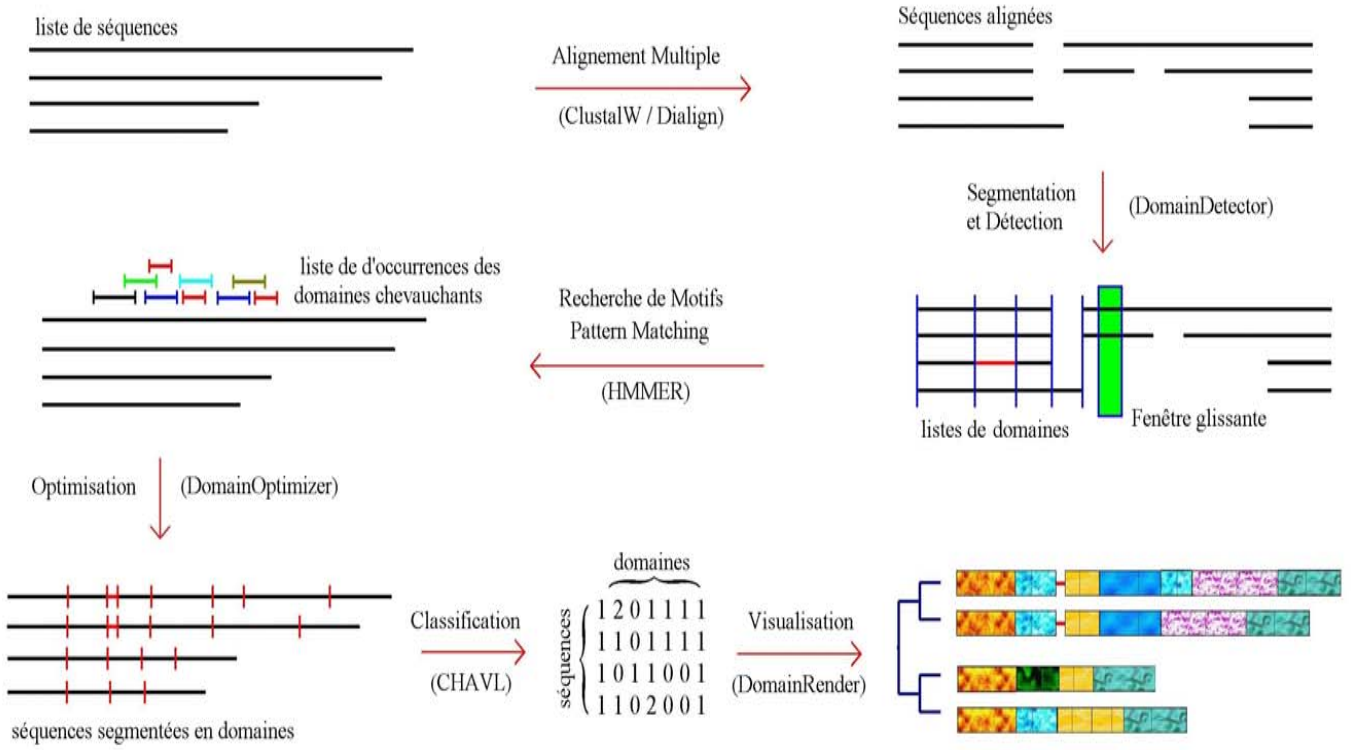
Couverture des séquences par des domaines : un problème d'optimisation.

Ens de séquences S (familles),

Ens de séquences D (domaines).

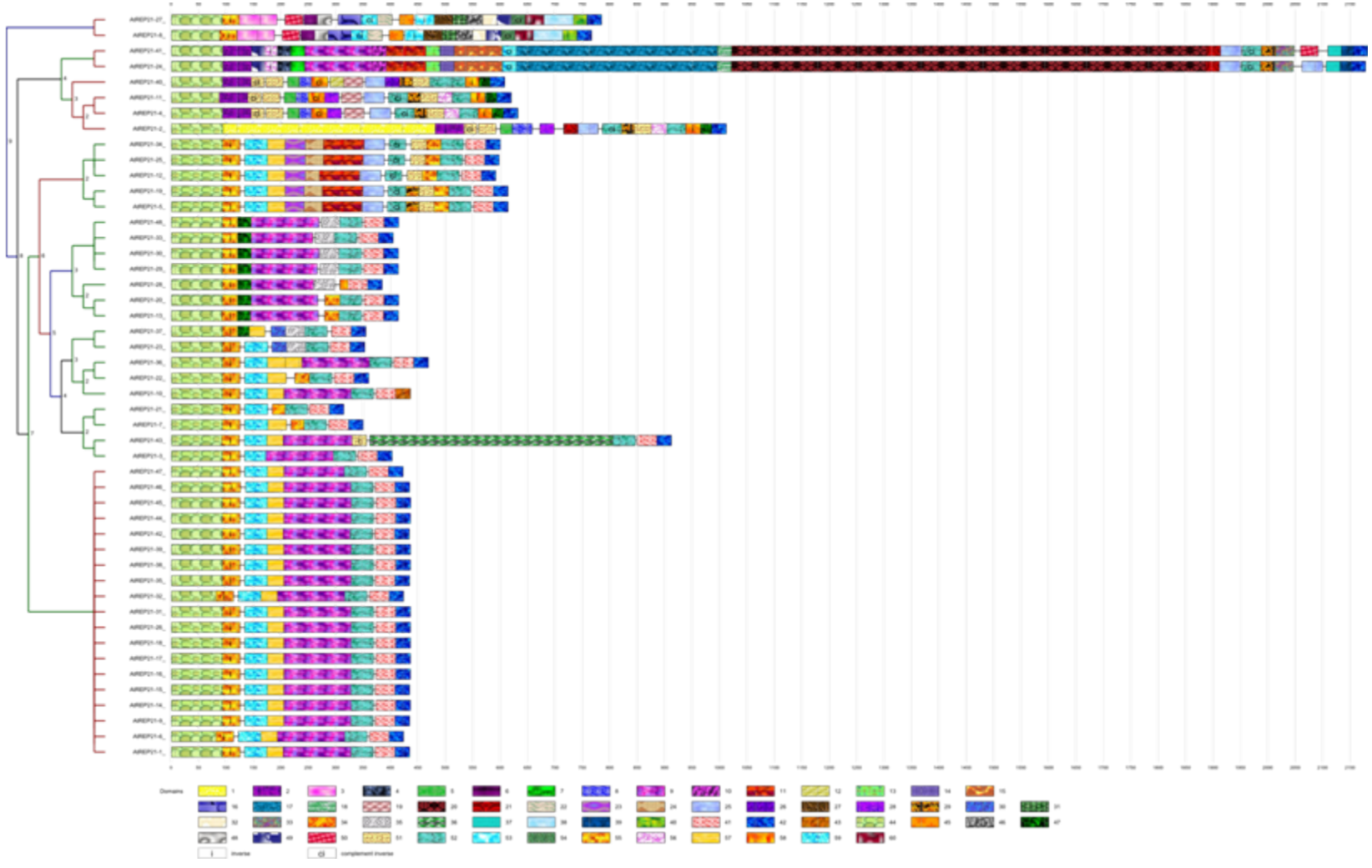
Trouver une couverture optimale de S par un sous-ensemble de D.

2006 Bioinformatics 22 (16)



Un résultat de DomainOrganizer

Famille ATrep3 chez A. thaliana

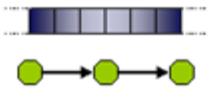
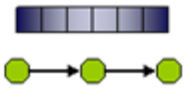
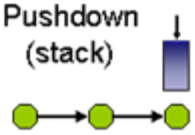
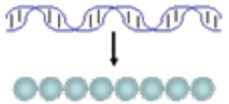


Analyse syntaxique

Modélisation syntaxique : Hiérarchie de Chomsky et Génomes

Constatation de D. Searls :

- des mécanismes élémentaires sur l'ADN comme la copie réclament une complexité descriptive et théorique élevée dans le cadre usuel.
- Utilisation de la variable de chaîne + morphisme comme objet syntaxique : SVG

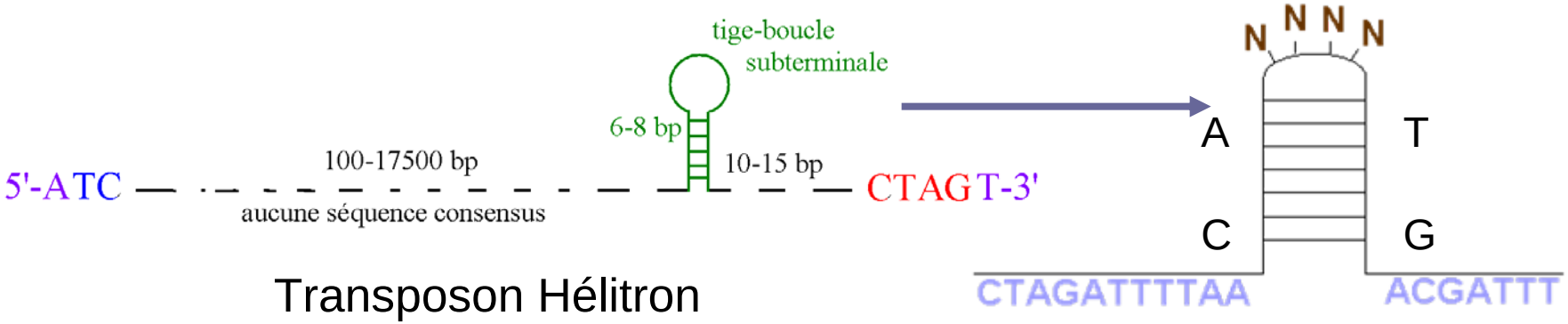
Languages	*Automaton*	*Grammar*	*Recognition*	*Dependency*	*Biology*
Recursively Enumerable	Turing Machine 	Unrestricted 	Undecidable	Arbitrary	Unknown
Context-Sensitive	Linear-Bounded 	Context-Sensitive 	NP-Complete 	Crossing 	Pseudoknots, etc. 
Context-Free	Pushdown (stack) 	Context-Free 	Polynomial 	Nested 	Orthodox 2^0 Structure 
Regular	Finite-State Machine 	Regular 	Linear 	Strictly Local 	Central Dogma 

Transposons : modèle syntaxique

- Analyseur syntaxique plein génome : STAN

  2005 Bioinformatics 21(24):4408 – 4410.

- *A.S. Valin, G. Ranchy, S. Tempel, P. Durand*



Transposon Hélitron

CTAGATTTTAA*:2* - X:[7] - x(4) - ~X*:3* - ACGATTT*:1*

- Motif
- Variable de chaîne
- Ensemble de bases quelconques
- Variable de chaîne complémentée

Erreurs

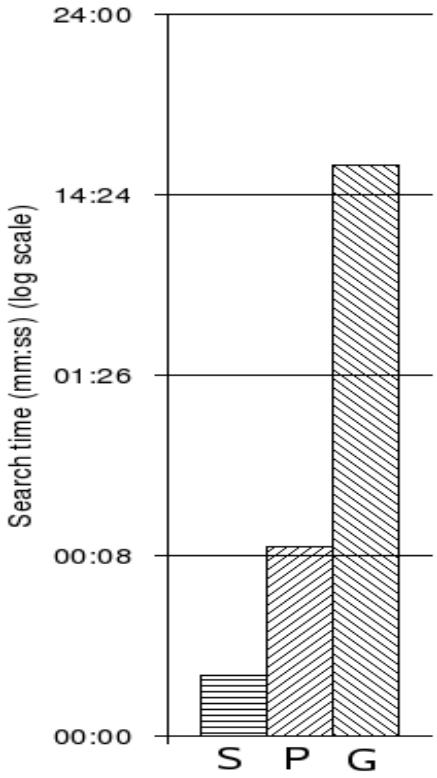
Hélitrons : modélisation et combinatoire

Génome cible : *A. thaliana* (131 Mo)
Résultats : 125 éléments trouvés (20sec)
SunFire 6800, 16 proc, 16 Go mém

- STAN.

- Caractérisation d'une famille complète de transposons, les hélitrons, à partir de leurs extrémités; Nouvelles familles découvertes.

  2007 Gene, 403(1-2):1299-1305.
  Thèse Sébastien Tempel 2007



S: STAN
P: PatScan
G: Genlang

Un nouveau langage de modélisation, Logol

- Face abstraite qui représente la chaîne d'origine et face concrète pour les différentes instances copies de cette chaîne d'origine.

  X::$1, X::$1 reconnaît le tandem repeat ACTA avec X=AA

- Analyse d'entités chevauchantes.

  X::$1; X::$1 reconnaît le tandem repeat ACAA avec X=AAA

- Intégration de contraintes d'optimisation

  (_, "ATG" :_address, (:#3, ?~stop)+, stop,_) : !@address
  reconnaît les zones codantes
  ATGATGTA**TGG**A**ATG**TAGCATGCGG**TAG**G

C. Belleannée, P. Durand
***Rapport de recherche INRIA**, number 6350, Nov. 2007*

Faces abstraites/concrètes : retrotransposons à LTR

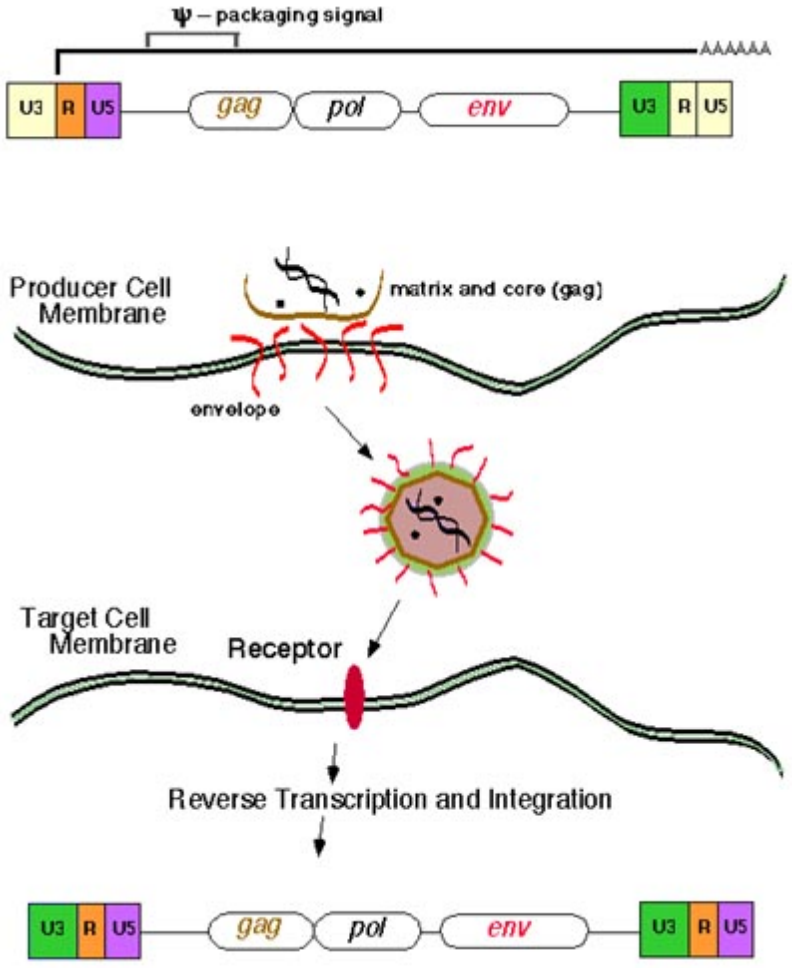
PERE

DR:[2..6],
(«tg», (U31,**R,U51**), «ca »),
[1..100], **pbs**, [1 000..20 000], ppt,
[1..100],
(«tg», (**U32**, **R:90%,** U52), «ca»),
DR

Paléontologie génomique

FILS

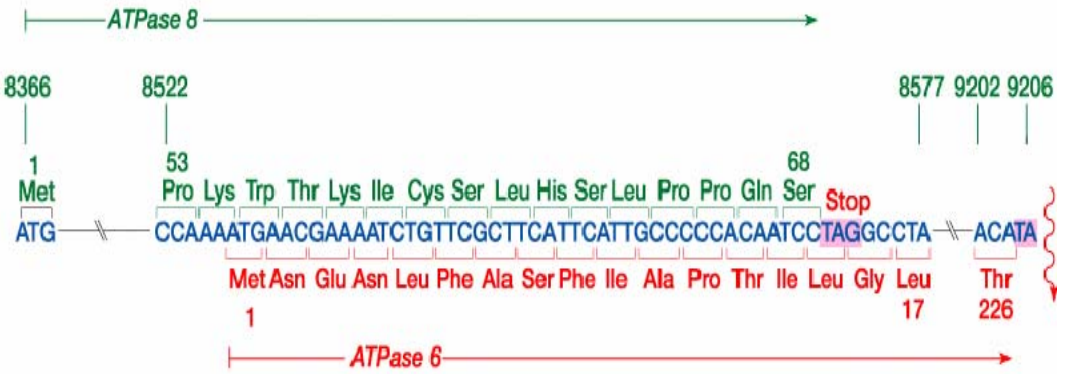
DR:[2..6],
(«tg»,**(U32,R,U51):95%,** «ca »),
[1..100],
pbs, [1 000..20 000], ppt, [1..100],
(«tg», **(U32,R,U51):95%,** «ca»),
DR



Projet ANR C. Belleannée

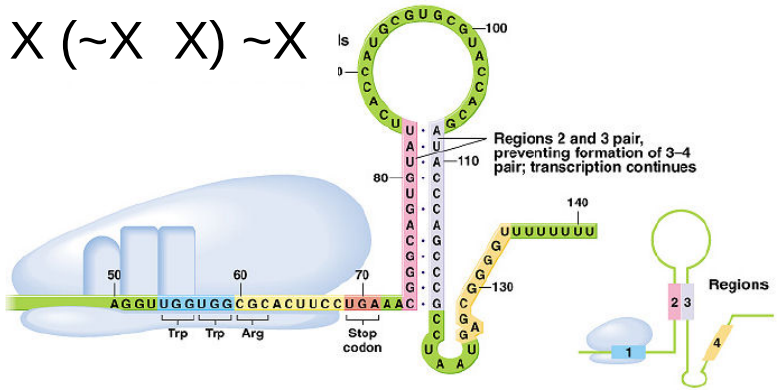
Chevauchement, ambiguïté et régulation

- Chevauchement sur le même brin de 2 gènes ATPase (8 / 6, homme).

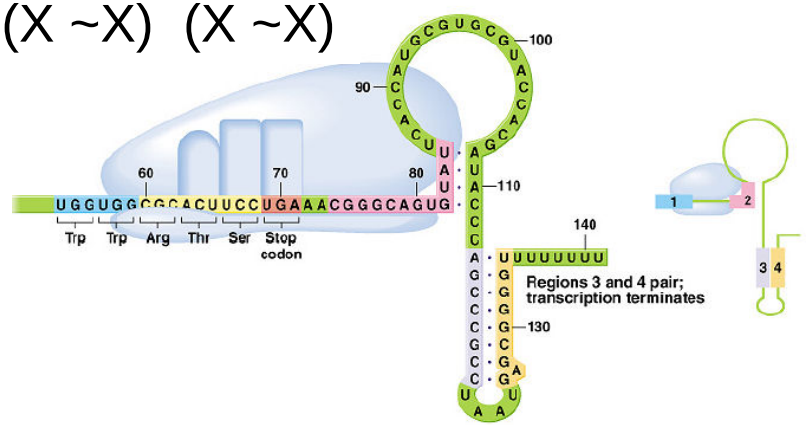


- Régulation par atténuation (bactéries)

X (~X X) ~X



(X ~X) (X ~X)

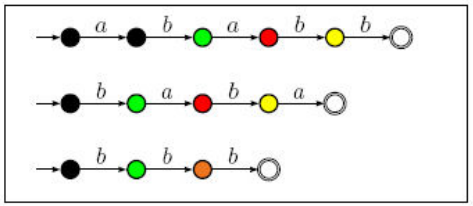


- Facteurs de transcriptions chevauchants (drosophile)

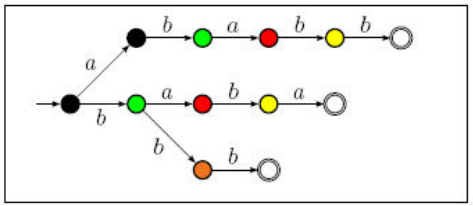
Bcd :SYWAATCC, Kr: WYAAYCCDDY, Séq:GTTAATCCGT

Inférence grammaticale

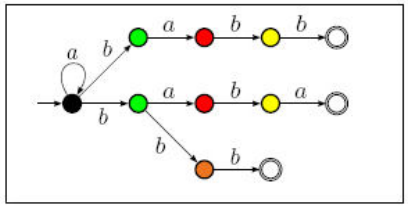
Inférence d'automates : principe



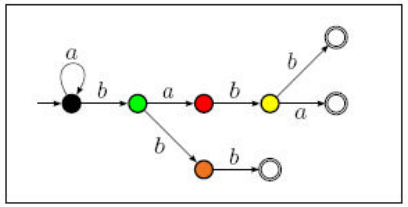
MCA



PTA

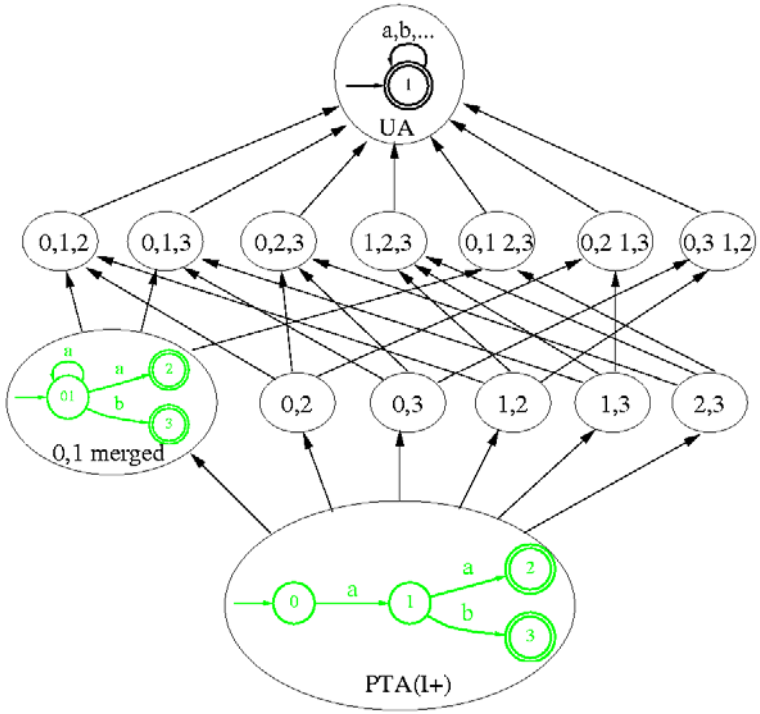


Automate après une fusion d'états.



Automate après une fusion déterministe.

- Etant donné un échantillon de séquences, on peut construire un treillis de l'ensemble des automates admissibles.
- Fusionner des états dans l'automate, c'est généraliser le langage reconnu.

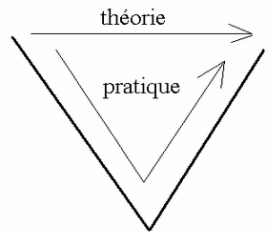


Quelques apports théoriques

- Inférence d'automates finis par gestion d'un ensemble de contraintes dynamiques.
  Thèse François Coste 2000

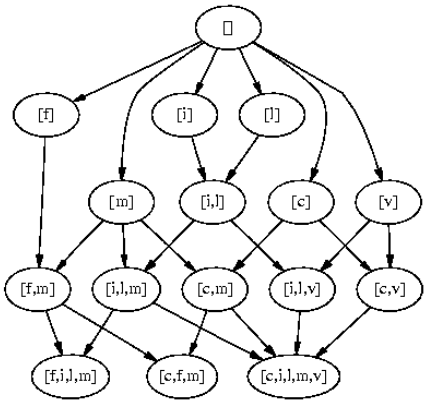
- Inférence d'automates non déterministes : traitement d'automates non ambigüs.
  Thèse Daniel Fredouille 2003

- Inférence de grammaires algébriques.
  Thèse Jean-Yves Giordano 1995

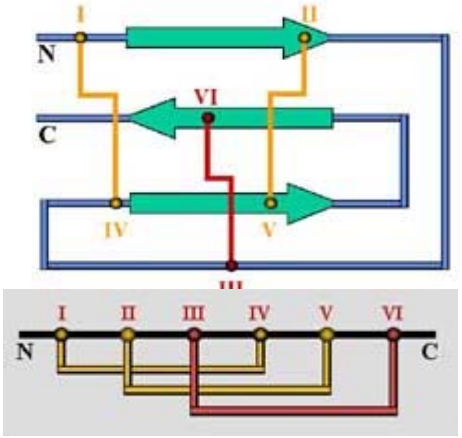


L'inférence grammaticale en pratique…

- Prise en compte de connaissances : alphabets ordonnés suivant les propriétés physico-chimiques des acides aminés
  SDTM *Thèse Aurélien Leroux 2005*



- Relations longue portée : prédiction de ponts disulfures dans les protéines
  Thèse Ingrid Jacquemin 2005



Un premier outil d'inférence publiquement disponible : Protomata

G. Kerbellec, F. Coste

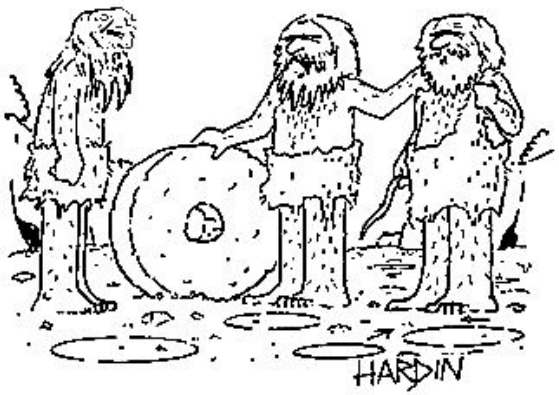
Un aperçu du renouvellement du bestiaire

Langages de Szilard : sur l'enchaînement des règles

1) S → c S
2) S → c S c S
3) S → a

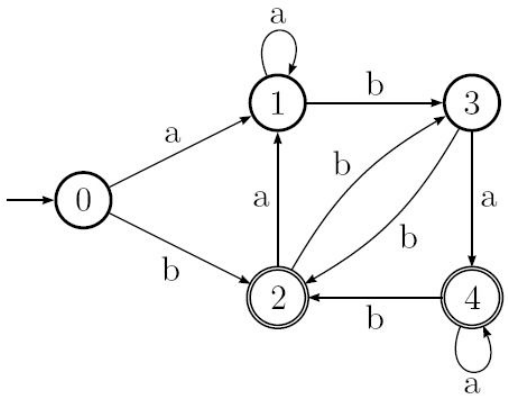
ccaccccad c(c(a)c) c (c(a)))

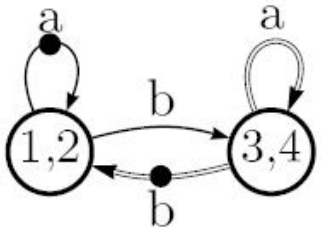
Séquence de contrôle 223113



"To be honest, I would have never invented the wheel if not for Urg's ground breaking theoretical work with the circle."

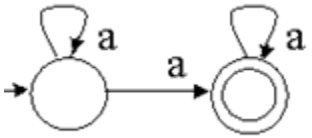
DFA

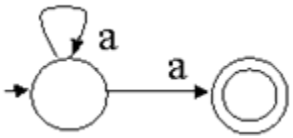


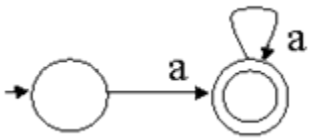


FTA

NFA → e^x → UFA → e^x → DFA







Trois exemples de parcours modélisation syntaxique/apprentissage/biologie

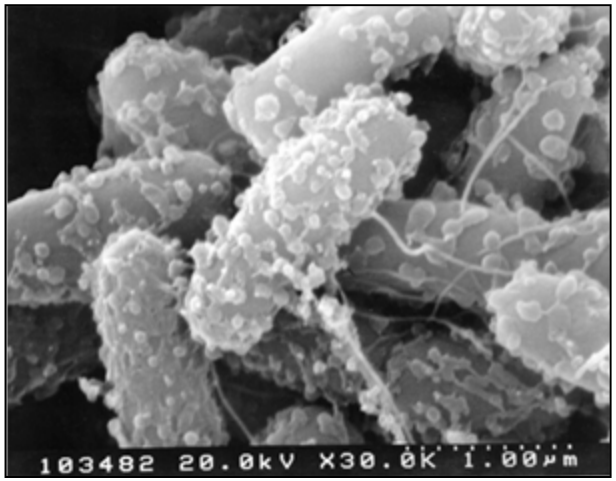
- Classification et apprentissage de signatures pour des canaux membranaires, changement de fonction par mutation dirigée : *Collaboration UMR 6026 C. Delamarche*

- Modélisation et découverte de défensines humaines : *Collaboration GermH INSERM C. Pineau*

- Modélisation et découverte de récepteurs olfactifs canins : *Collaboration UMR 6061 CNRS F. Galibert*

Recherche de β-défensines humaines

- Les défensines sont une alternative potentielle aux antibiotiques dans le contexte de la multiplication des micro-organismes multi-résistants.
- 30 nouvelles défensines découvertes, 26 brevetées.
- *GermH Inserm*
  Y. Bastide, Y. Mescam, F. Bourgeon, C. Pineau



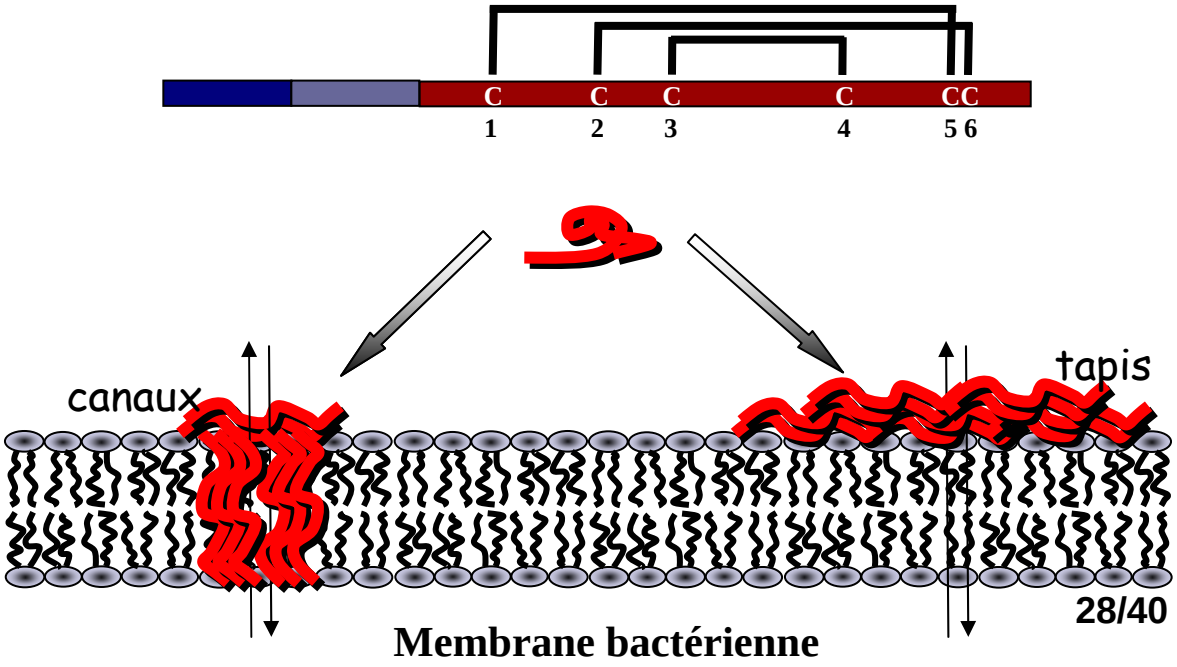


Schéma d'expérimentation :

Banques

Extraction

+ β-défensines

— α-défensines

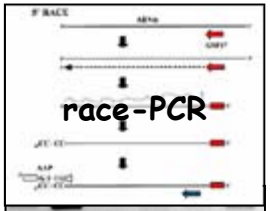
chemokines

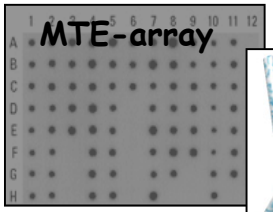
Apprentissage
signature caractéristique

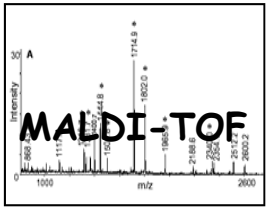
C-x(4,6)-T-X(0,5)-C-x(7,8)-C-x(4)-[DEFG]-x(2)-[SVP]-[YEAST]-C-x(7,9)-C-C

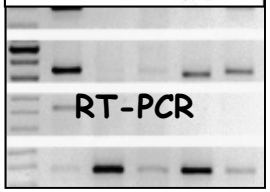
Recherche dans les génomes humain et murin

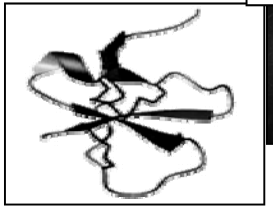
séquences candidates
> CLQNGGFCLRSSCPSHTKLQGTCKPDKPNCC
> CLTKGGVCWGPCTGGFRQIGTCGLPRVRCC
> CLQHGGFCLRSSCPSNTKLQGTCKPDKPNCC
> CHTNGGYCVRAICPPSARRPGSCFPEKNPCC
>…



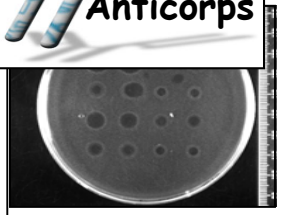




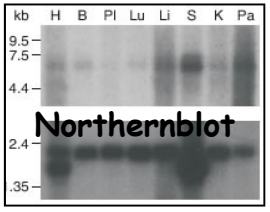




peptides de synthèse



Tests biologiques



Validation in vitro, in vivo

Validation in silico

Thèse F. Bourgeon 2003

gene name	chr	occurrence	charge	C-spacing	transcript	Mouse hit	Score	TrackP.	HMM	Other
DEFB104	8	CGYGTARCRKKCRSQEYRIGRCPNTYACC	6	6-3-9-5	✓	*	7	✓	✓	✓

Découverte de gènes de récepteurs olfactifs

UMR CNRS 6061 P. Quignon, P. Lavigne, F. Galibert.
M. Giraud, D. Lavenier, E. Morin, E. Retout, A. S. Valin,
Genome Biol 6, no. 10 (2005). + JOBIM 2005, 81-87

Séquençage

(2004)

36 M STS
36 Gbp

Assemblage

6-12 mois

Brouillon
d'assemblage
Juillet 2004

PC : 2 jours
R-Disk : 15 minutes

Découverte & Appariement
de 5 motifs dégénérés

63745 STS
60 Mbp

Assemblage ciblé

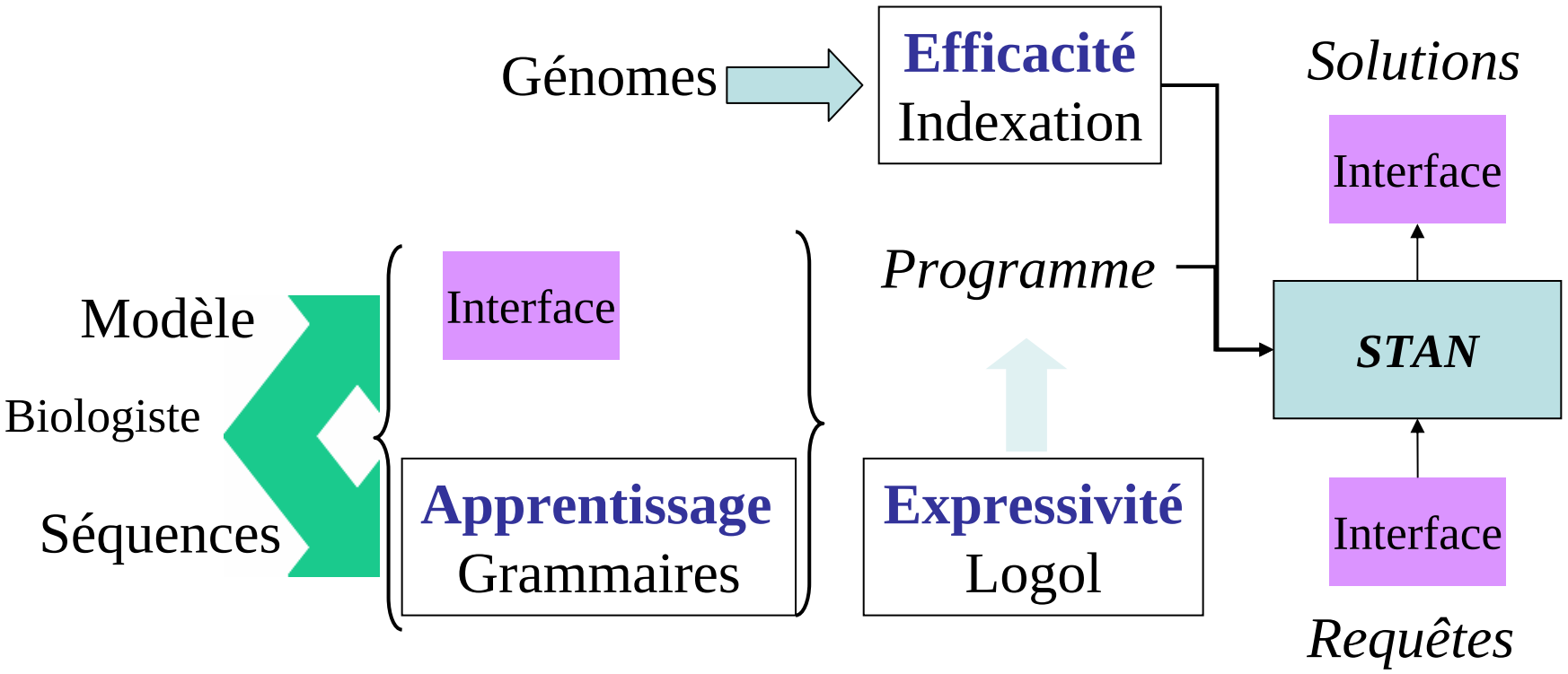
10 heures

Appariement
de 5 motifs
dégénérés

5568 contigs
1122 gènes OR

Expertise,
Prédiction de gènes

En Résumé…


Génomes
Efficacité
Indexation
Solutions
Interface
Programme
Modèle
Biologiste
Interface
STAN
Séquences
Apprentissage
Grammaires
Expressivité
Logol
Interface
Requêtes

Pour conclure

Perspectives : vers un nouveau challenge …

- Un constat: Biologie très tributaire de la poussée technologique (exemple puces à ADN). **Le bioinformaticien doit être un acteur des nouvelles questions scientifiques** accessibles;

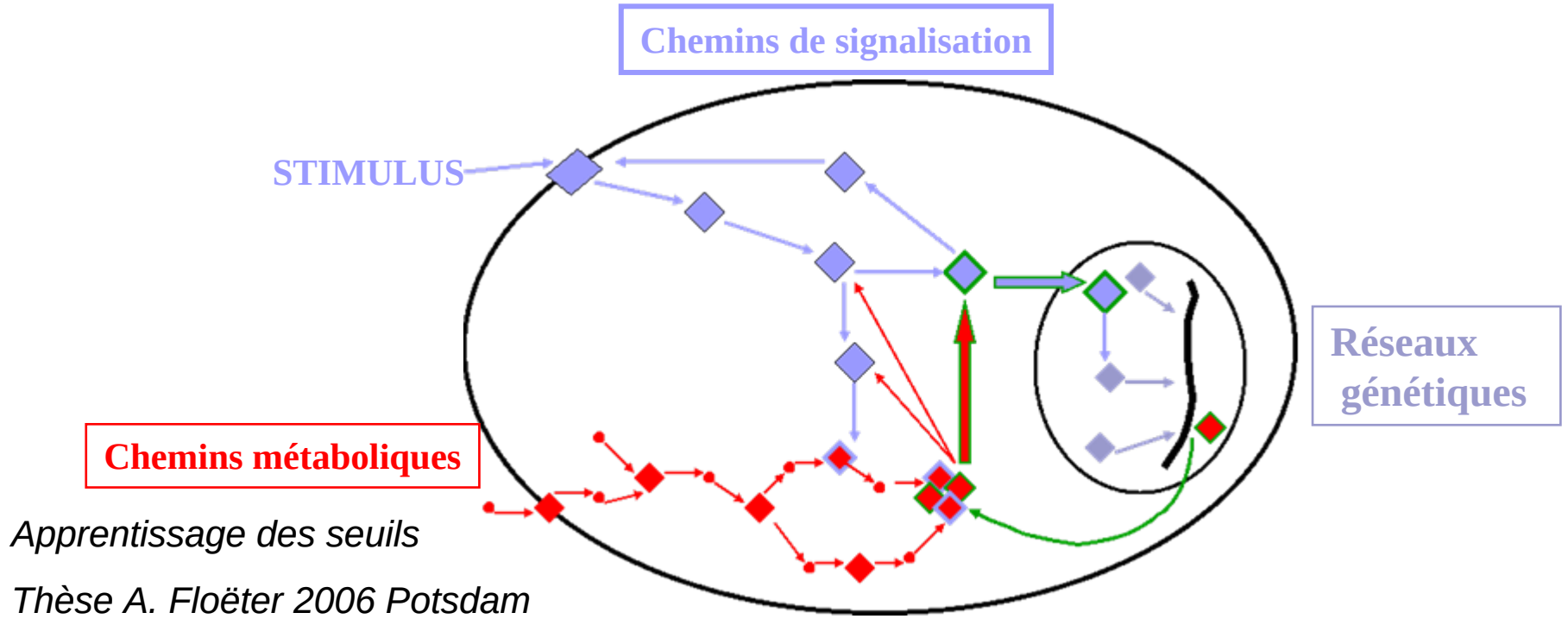
- Effort massif d'intégration des données de la communauté bio/bioinfo, mais peu d'aide au raisonnement (exception F. Fages, Rocquencourt);

- Expérimentation : Laboratoires sur puce en plein essor. Les problèmes principaux restent la généricité et l'intégration;

- Réseau d'expériences en modélisation (**ACI MathResoGen, thèse cotutelle Potsdam),** en μ/nanotechs (ENS Cachan, IETR), logique (Imperial College, Potsdam) et expérimentation cellules (Inserm).

Réseaux d'interactions dans une cellule

- Complexité de la connexion de 3 réseaux entre eux



Apprentissage des seuils

Thèse A. Floëter 2006 Potsdam

- Bases de connaissances instables mais nombreuses
  - Kegg, Metacyc, Reactome, HMDB, Metacrop, Enzyme, GO, …
  - CSNDB, Transpath, Transfac, Genenet, Prodoric,…

Expérimentations en Biologie, un compromis

+Coût

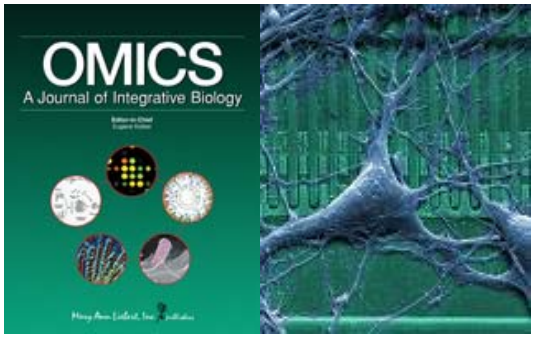
+Validité

+ Repro-ductibilité

+ Grande échelle

+Réutili-sabilité

- In vivo : experimentation sur animaux
- In vitro : experimentation sur cellules
- In silico : experimentation sur modèles

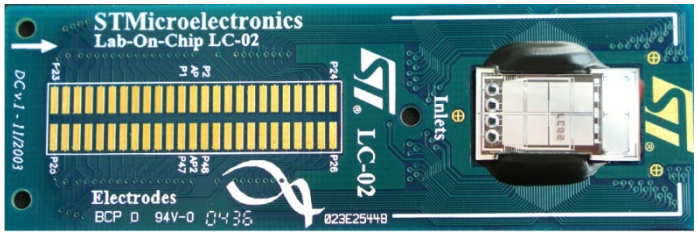


- Lab-on-chips pour la “cellomics”: coupler in vitro et in silico

Lab on chips (µ Total Analysis Systems -µTAS)

- Identification de souches bactériennes
  CEA Grenoble

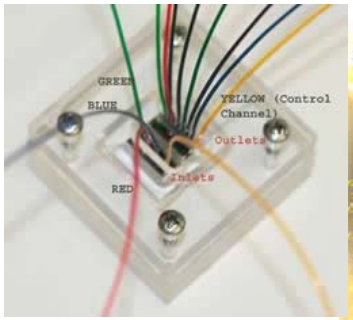
Fouqué B, Brachet AG, Marcel F, Dupont R, Delapierre G, Fischetti A, Jalava J and Chatelain F MicroTAS 2005, October 9-13th, Boston, USA.

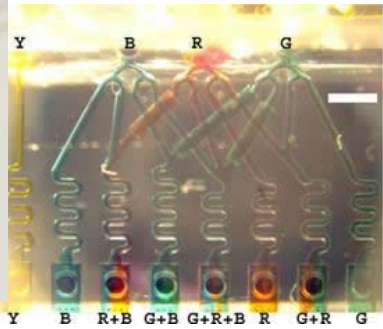


- Surveillance de l'eau de mer

  UCLA Los Angeles

S. Gawad, K. Cheung, U. Seger, A. Bertsch, P. Renaud, Lab on a Chip 4, 241 (2004)



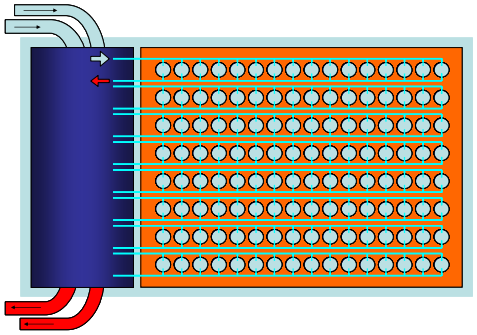


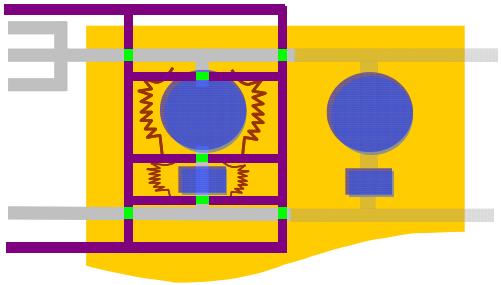
- *Détection de protéines surproduites en cas d'attaque cardiaque (creatinine kinase, myoglobin, troponin I, et BNP) en 15 minutes*
  Triage Cardio ProfilER

Le projet Basic Lab :

- Laboratoire intelligent sur puce sur lamelle standard.
- Automatiser la boucle : Expérimentation intensive, observation et acquisition de connaissance.
- Micro-fluidique, micro-capteurs et module de raisonnement complet (déduction, apprentissage par induction et planification).

Précurseur: Robot Scientist R. King, Computational Biology Group, University of Wales, Aberystwyth, UK. ToxDrop Grenoble





Expérimentations sur réseaux biologiques : un cadre idéal pour le raisonnement automatique

- K= connaissance a priori sur le domaine;
- I= conditions initiales impsées par le dispositif expérimental;
- O= observations;
- H= classe d' hypothèses alternatives.

- Déduction :
  - $K \cup I$ est-il consistent par rapport à O ?
  - Quelles sont les conséquences possibles des conditions initiales I ?
- Abduction :
  - Quel $S_K \subseteq K$ explique au mieux la consistance de $S_K \cup I$ / O ?
- Induction :
  - Existe-t-il une nouvelle hypothèse $h \in H$ expliquant O ?
    ($K \cup I$ ne permet pas de déduire O, $K \cup I + H$ permet de déduire O)
- Planification :
  - Quelle expérience permet de discriminer au mieux parmi les hypothèses concurrentes H ?

Acteurs du projet Basic Lab

- Coordination, Modélisation, Architecture :
  Symbiose
  (T. Hénin, A. Siegel, D. Lavenier, P. Veber, C. Vargas);
- Logique, Résolution combinatoire :
  Potsdam Universität
  (T. Schaub, M.)
- Microfluidique :
  Biomis ENS Cachan Bruz *(C. Jullien, D. Grenier, F. Razan)*
- Capteurs, électronique :
  IETR Rennes *(S. Crand, T. Brahim)*
- Expérimentation cellules humaines :
  INSERM Rennes (*C. Guillouzo, R. Le Guével)*
- Expérimentation bactéries :
  Duals UMR 6026 (*F. Hübler, S. Avner, D. Thybert)*

Merci !

- Au jury
- A toute l’équipe Symbiose
- A l’ancienne équipe Aïda
- A toutes les personnes qui ont supporté mes conseils avisés jusqu’à leur soutenance de thèse : Catherine, Raoul, Francis, Jean-Yves, Robin, François, Daniel, Aurélien, Ingrid, André, Sébastien

Et une exhortation spécifique pour les enseignants : faites vous plaisir en recherche, vous ferez le bonheur de vos étudiants !