



Annotation des génomes complets

Seminaire IN'Tech
bioinformatics : from genomics and post-
genomic data to biological knowledge

Yves Vandembrouck

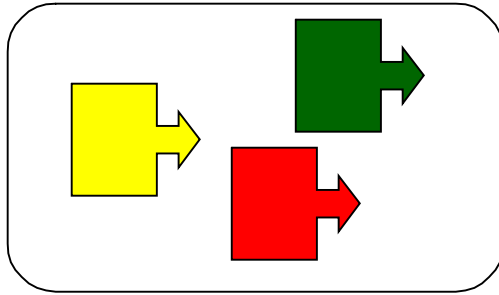


23 octobre 2003, Lyon

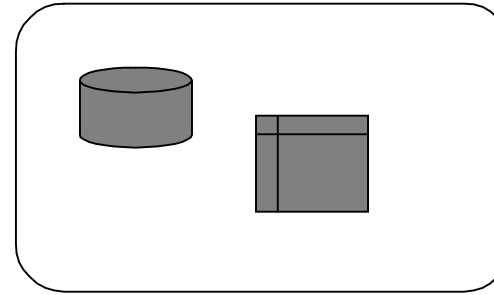
Environnements d'annotation (1)



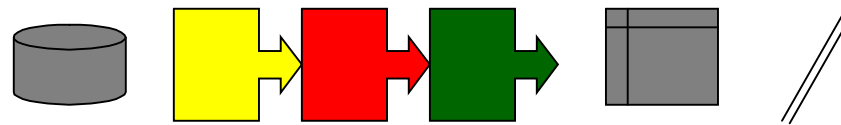
outils



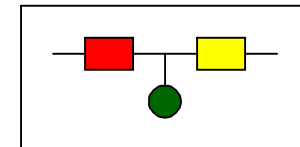
données



- « pipe-line »



Interface

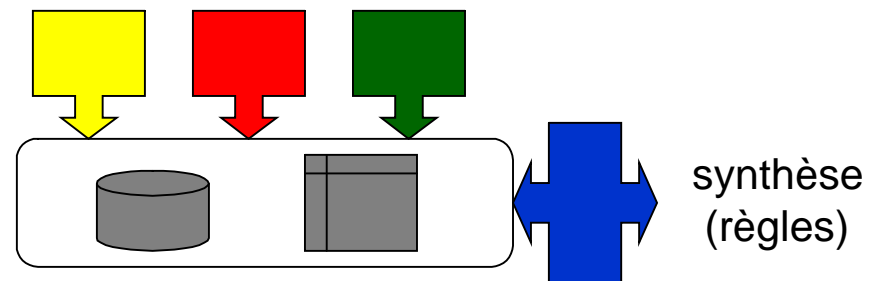


- gros débits
- programmation
- interfaces « passives » (consultation)

ex: *Pedant*, *GeneQuiz*, *BioFacet*

- annoteurs automatiques

Magpie (Gaasterland et al.)

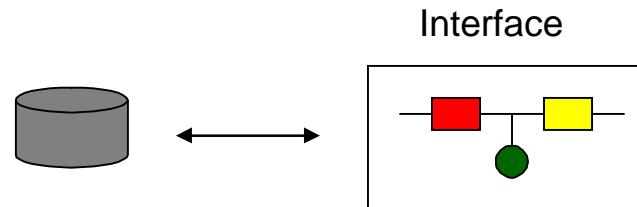


Environnements d'annotation (2)

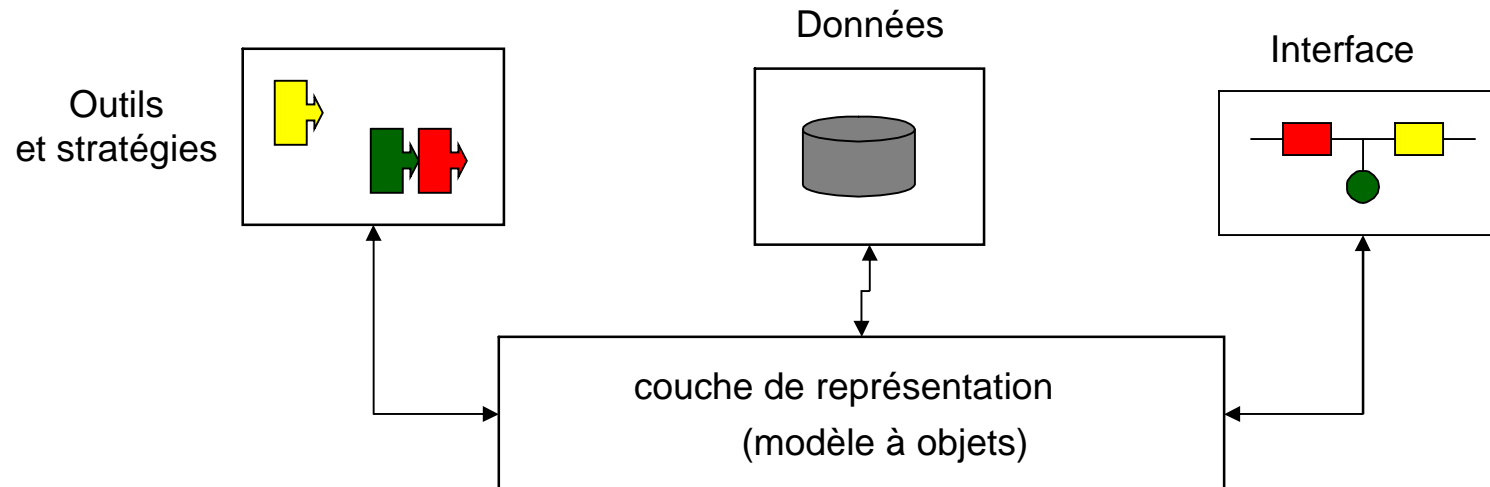


- éditeurs d'annotations

Genotator (Harris et al.)
Artemis (Rutherford et al.)

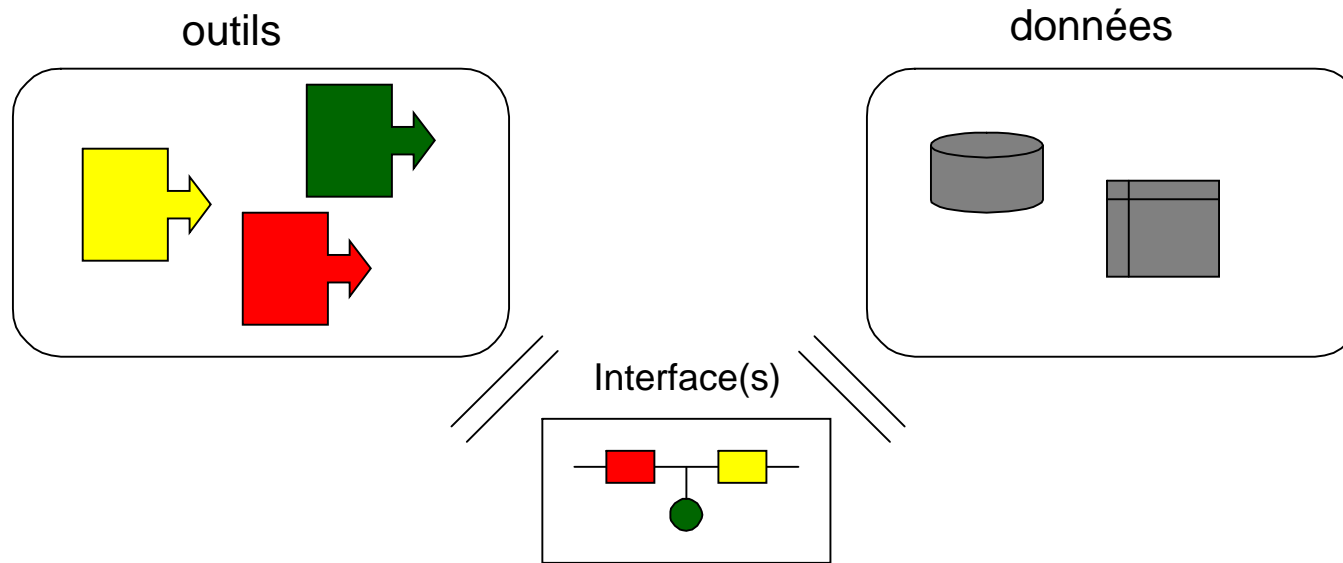


- environnements interactifs



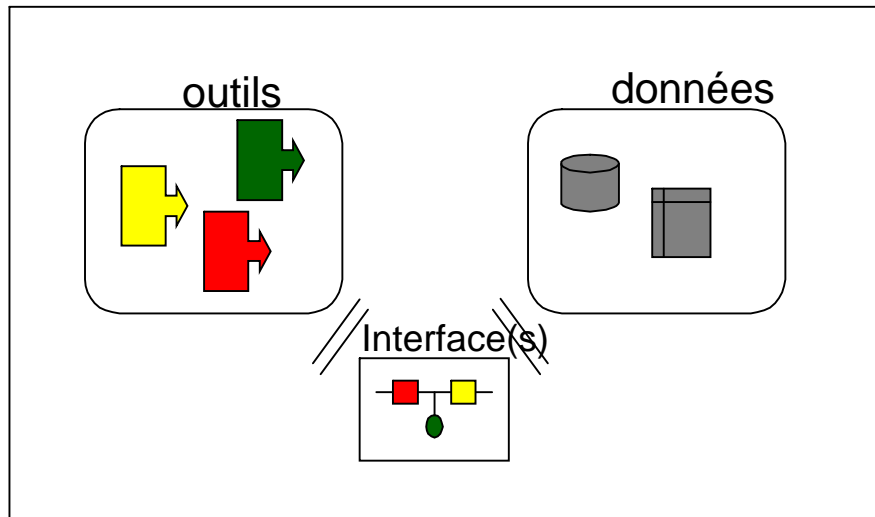
Imagene (Médigue et al. 99)

Environnement d'annotation : les objectifs

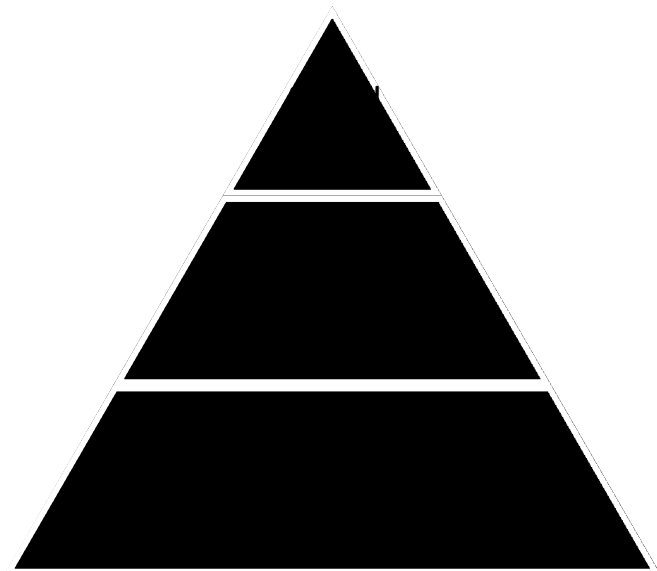


1. Confronter différentes sources de données : Hétérogènes, distribuées,
2. Appliquer des méthodes d'analyse dédiées : Stratégies
3. Favoriser le travail de l'expert-annotateur : Interfaces utilisateur (contrôles des paramètres, visualisation des résultats, édition des annotations)

L'environnement GenoStar



GenoStar : Système de représentation des connaissances *

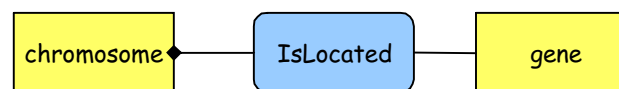


Entity / relation
Classes/Associations

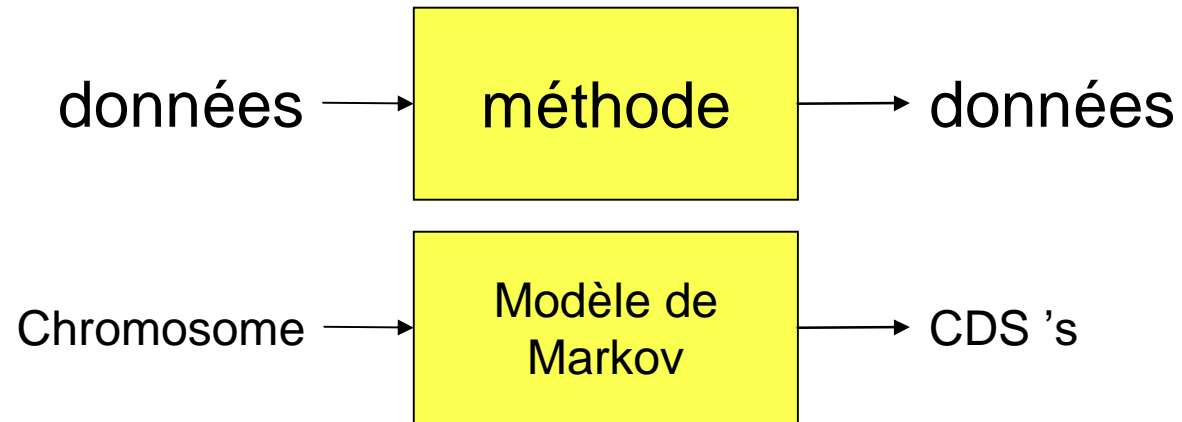
UML



Object Oriented
Representation System



GenoStar : Méthodes (1)



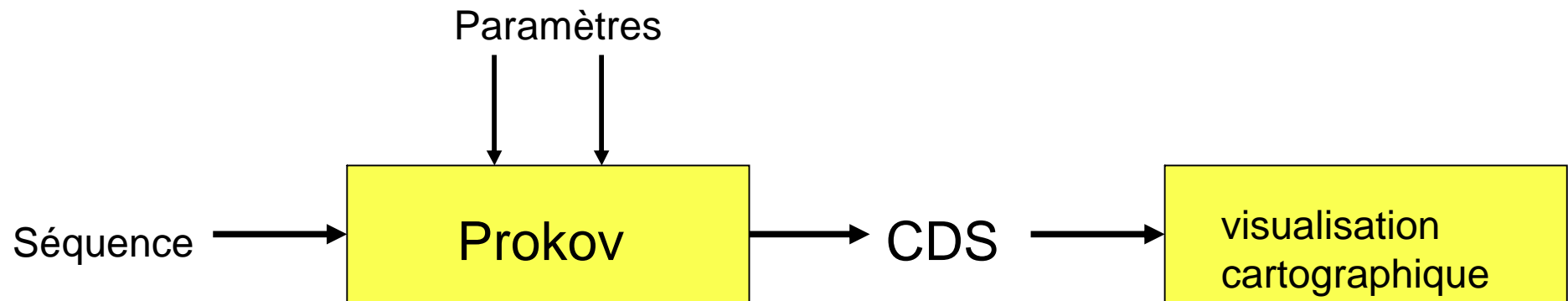
- représenter et organiser les méthodes comme des entités

Modélisation Connaissance méthodologique

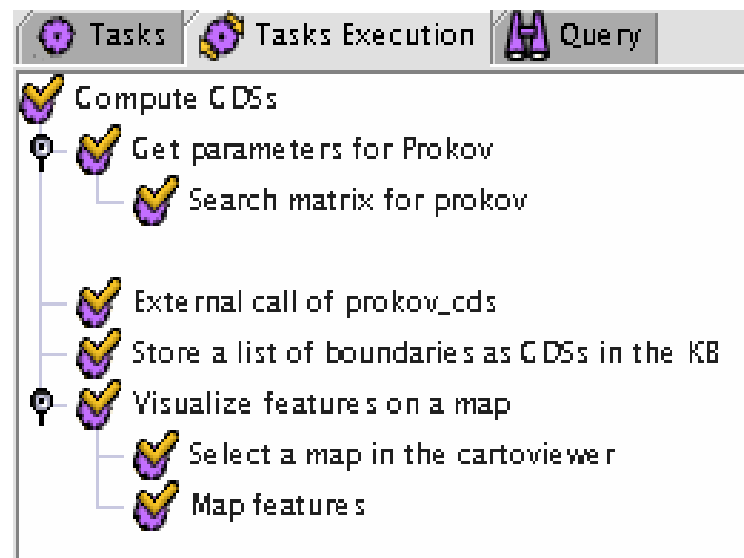
- contrôler l'exécution (quelle méthode pour quelle donnée)

GenoStar : Méthodes (2)

- Des tâches paramétrables enchaînant ces méthodes sur les données



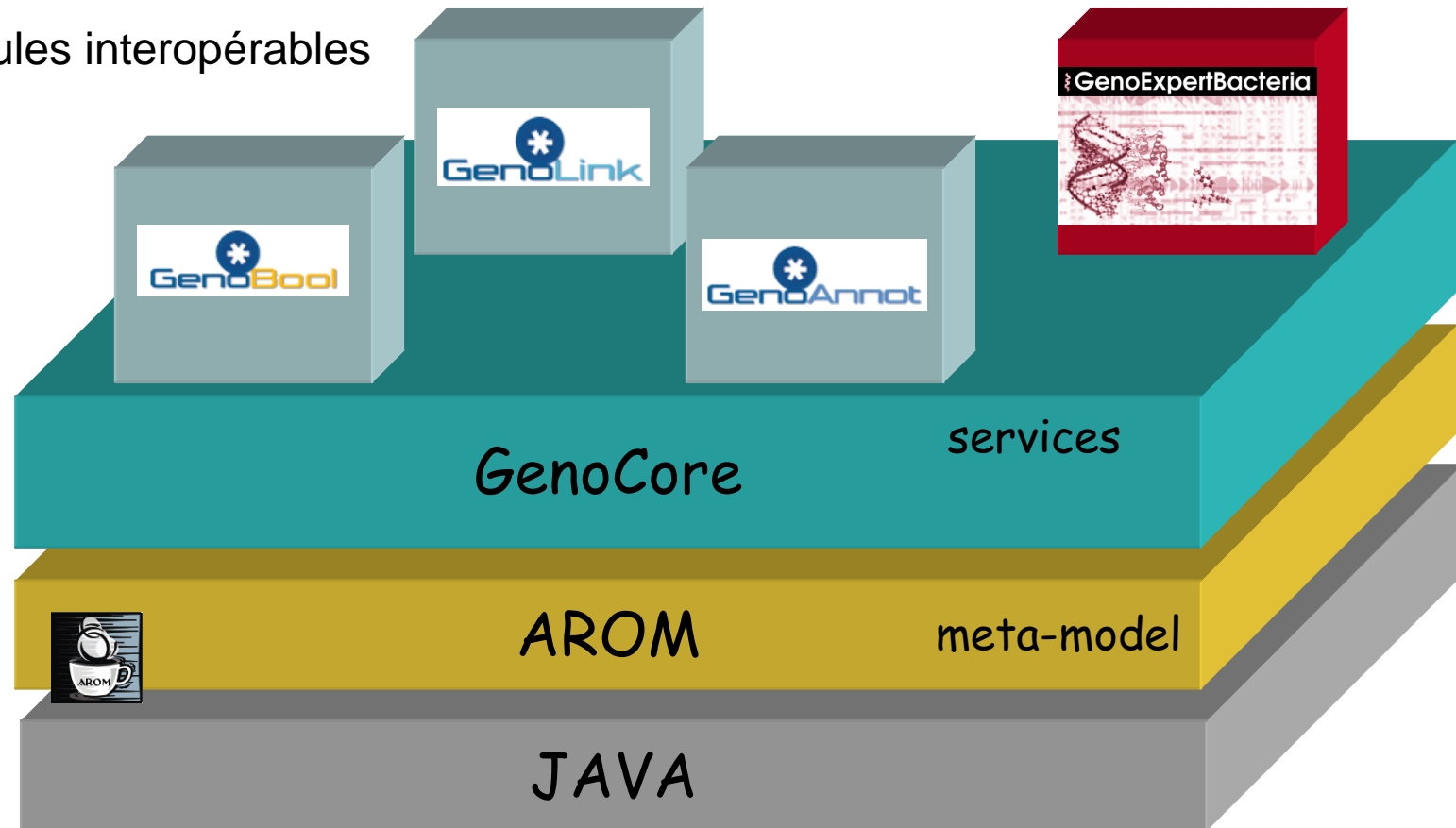
Décomposition
logique



GenoStar : Architecture



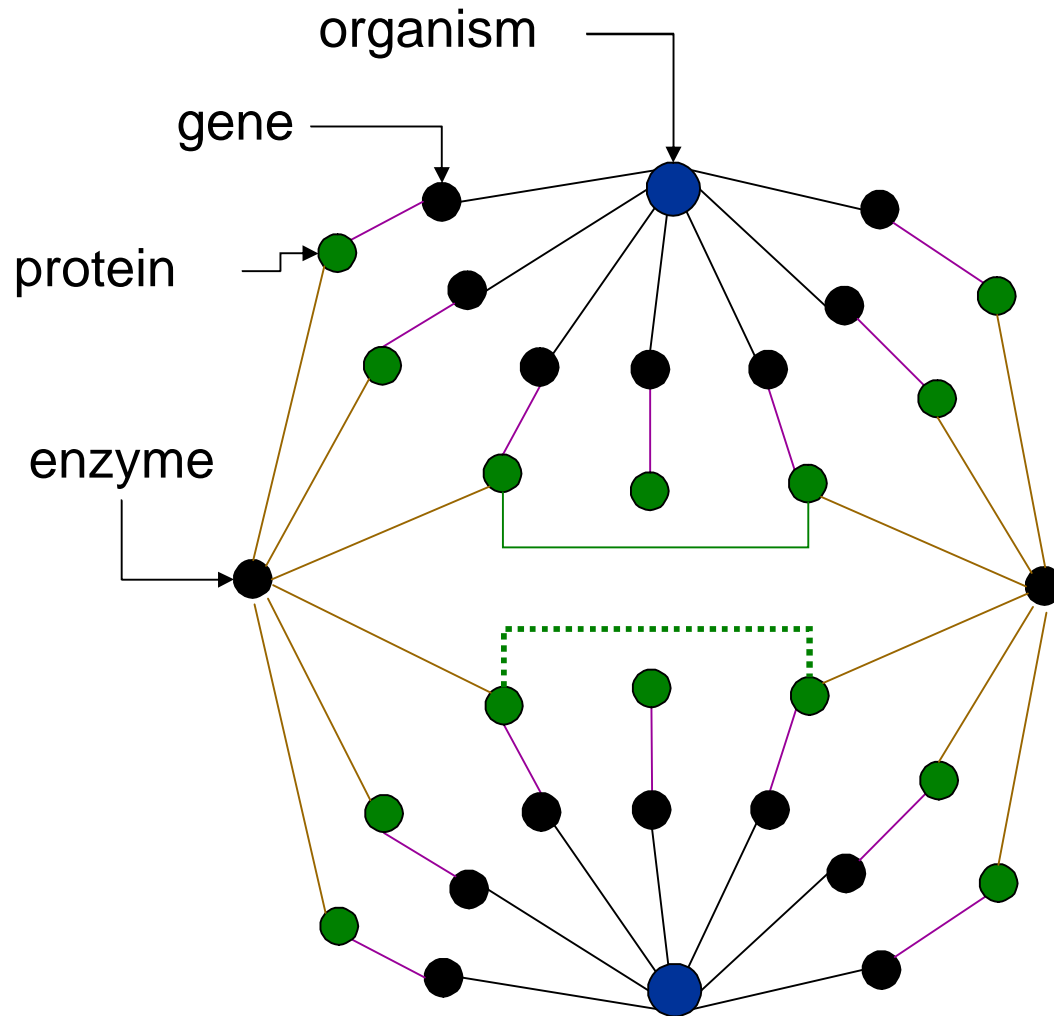
Des modules interopérables



autour d'un noyau qui assure :

- la gestion des données et des connaissances
 - l'enchaînement de l'exécution des méthodes d'analyse

GenoStar : Environnement de génomique exploratoire



Annotation de séquences

- Identifier les objets



Annotation fonctionnelle

- Explorer le voisinage



GenoExpertBacteria



Fouille de données

- Découvrir de nouveaux objets / relations



GenoAnnot : annotation séquence brute (1)



- Gestion des données et des tâches :

Contig ~ 45 kb (AF305077, GenBank) *Anaplasma marginale*

The screenshot shows the GenoAnnot software interface. The main window is titled 'AnmaGSC.wsp'. The 'Collections' pane on the left shows a tree structure with 'Object Collections' expanded, containing 'User Sequences', 'Organisms', and 'IMPORT_GENBANK_1'. The 'Objects [2]' pane shows 'ORG_Anmar' and 'SEQ_AF305077'. The 'Editeur de propriétés' window is open, showing the instance 'ORG_Anmar of GA_Organism'. The 'Tasks Execution' panel on the right shows a list of tasks with numbered callouts (1-7) indicating the current step in the workflow.

Editeur de propriétés

Instance ORG_Anmar of GA_Organism

Variable	Value
history	
scientificName	Anaplasma marginale
commonNames	Set Anaplasma marginale <input type="button" value="Edit"/>
strain	St. Marie
taxonomy	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; Anaplasmataceae; Anaplasma
taxon	770
nbChromosomes	1
ploidy	1
geneticCode	11

Tasks Execution

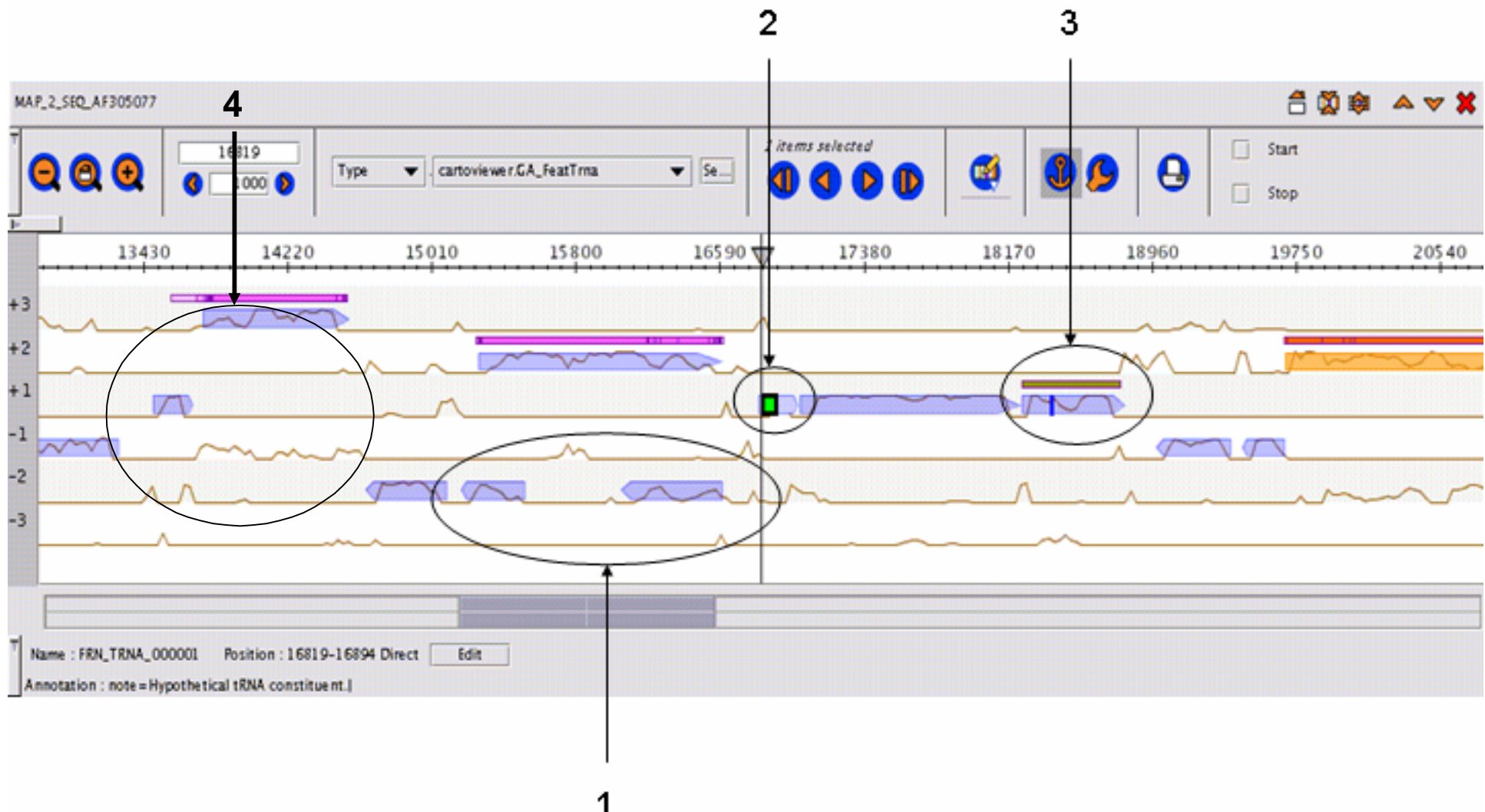
- Import_Export
 - Import fasta sequence
 - Import EMBL
 - Import GenBank **1**
 - Export in EMBL format
- Data_Manipulation
- Visualization
 - Display features from a collection
 - Display features associated with a sequence
- CDS_Prediction
 - Prokov
 - Build Markov matrix from sequence **2**
 - Build Markov matrix from a collection of CDSs **3**
 - Compute CDSs **4**
 - Compute coding curves **4**
- RNA_Prediction
 - Compute tRNAs **5**
- Signal_Prediction
 - Compute Rho-independent terminators **6**
- DB_Screening
 - BlastP on CDSs collection
 - BlastP on one CDS
 - BlastX **7**
- Administration

Display features from a collection
input: set-of-GA_Feature
documentation:

GenoAnnot : annotation séquence brute (2)



- Interface cartographique : visualisation des résultats des méthodes



GenoAnnot : annotation séquence brute (3)



- Interface cartographique : visualisation des résultats des méthodes
- Editeur de propriétés : statut des features

The screenshot displays the GenoAnnot software interface. On the left is a property editor for an instance of GA_FeatGds. The 'Annotations' tab is active, showing a table of variables and their values. A dropdown menu for 'qualitativeConfidence' is open, showing options: 'unvalued', 'probable', 'putative', 'sure', and 'wrong'. The 'probable' option is selected. Below this, the 'On Sequence' dropdown is set to 'SEQ_AF305077'. At the bottom, another table shows feature coordinates: start position 18250, end position 18807, and length 558, with a 'Direct' orientation.

On the right is a genomic map showing sequence data with various annotations. A scale at the top indicates positions from 17380 to 20540. A blue bar represents a feature, with a green square at its start (position 18250) and a red square at its end (position 18807). Two arrows labeled '2' and '3' point to the start and end of this feature, respectively. A double-headed arrow connects these two points, indicating the feature's extent. The map also shows other tracks, including a signal track and a track with orange and red bars.

GenoAnnot : annotation séquence brute (2)



Utilisation de blastp et assignation fonctionnelle primaire

The screenshot displays the GenoAnnot interface. The top window shows a BLAST search for sequence AM_000053. The search parameters include the query sequence (1342 letters) and the database (transmbSP). The search results show a significant hit with a score of 11.0. The annotation editor window, titled "Editeur d'annotation", is open over the search results. It shows the product name "probable surface protein precursor" and a note field. The interface includes a sequence viewer with a scale from 0 to 4005, a BLAST alignment view, and a table of search results. The table lists the following sequences:

Accession	Gene	Function	Citation	Note	Standard Name	EC Number
gi 4894576 gb AAD32553.1 AF117273	spl066818 Y539_AQUAE	[285 aa.] Hyp				
splQ50585 MMLC_MYCTU	[1146 aa.] Pu					
splQ9CKC8 ZIPA_PASMU	[323 aa.] Ce1					

The annotation editor window also includes a "Re set Card" button and a "Validate" button. The search results table includes a "note" column with the text "probable surface protein precursor".

GenoAnnot : reannotation de génome (1)



Génome de *Borrelia b.* : fragment des premiers 100 kb [Fraser et al. 97]

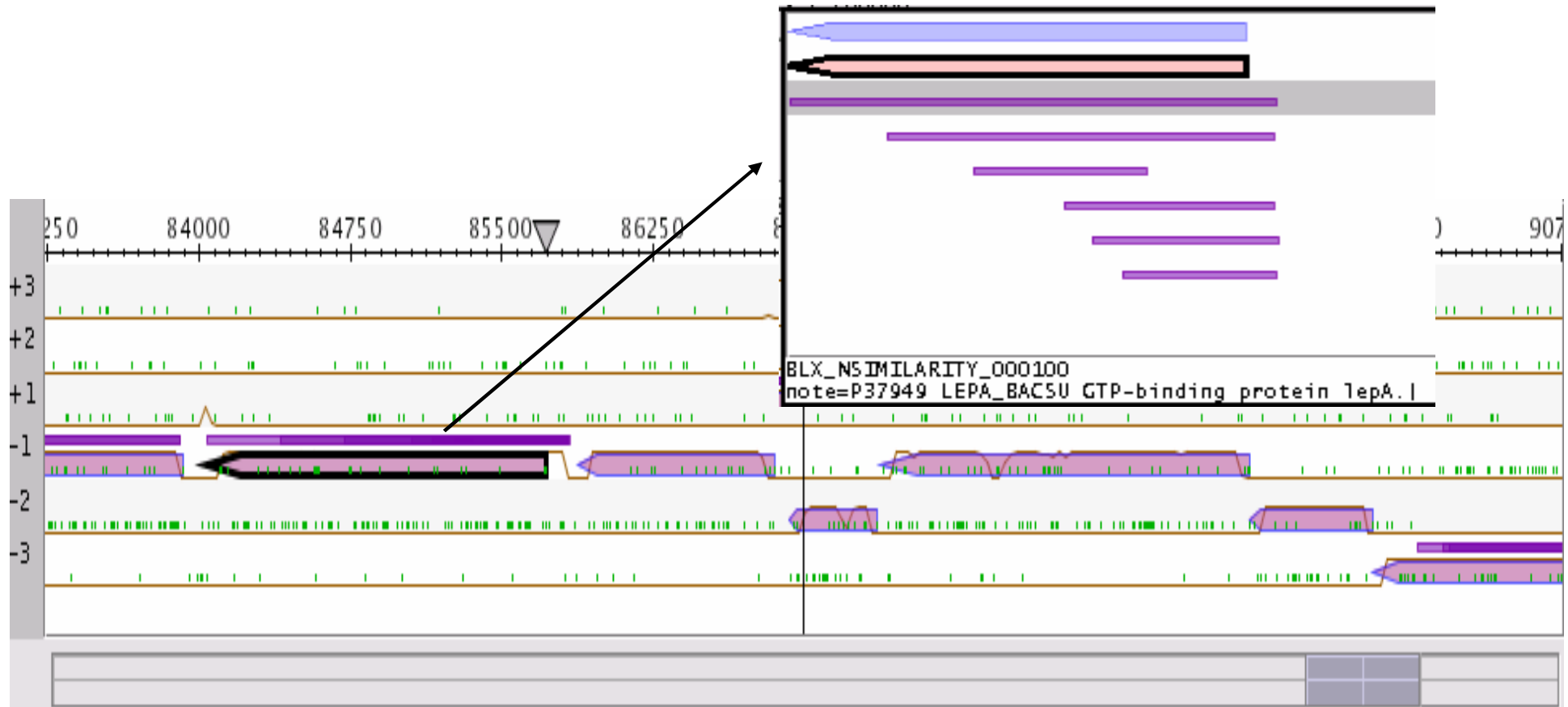
The screenshot displays the GenoAnnot software interface. The top menu bar includes 'File', 'Collection', 'Object', 'Query', 'Tasks', and 'Window'. Below the menu is a toolbar with various icons. The main workspace is divided into three panels:

- Collections:** A tree view showing 'Object Collections' with sub-items: 'User Sequences', 'Organisms', 'IMPORT_GENBANK_1', 'FEATURE_SEQ_AE000783', and 'PROKOV_CDS_1'. A circle highlights the 'User Sequences' folder.
- Objects [2]:** A list of objects: 'SEQ_AE000783' and 'SEQ_AE000783_1_100000'. An arrow points from the 'User Sequences' folder in the Collections panel to this object list.
- Tasks:** A tree view of tasks with sub-items:
 - Import_Export
 - Import fasta sequence
 - Import EMBL
 - Import GenBank (1)
 - Export in EMBL format
 - Data_Manipulation
 - Cut a fragment (2)
 - Map features
 - Visualization
 - Display features from a collection
 - Display features associated with a sequence (3)
 - CDS_Prediction
 - Prokov
 - Build Markov matrix from sequence
 - Build Markov matrix from a collection of CDSs
 - Compute CDSs
 - Compute coding curves (4)
 - RNA_Prediction
 - Signal_Prediction
 - DB_Screening
 - BlastP on CDSs collection
 - BlastP on one CDS
 - BlastX (5)
 - Administration

GenoAnnot : reannotation de génome (2)



- Combinaison de résultats de plusieurs méthodes : i.e correction de start



Interoperabilité GenoAnnot -> GenoBool (1)



- Importation des CDS dans GenoBool
- Recherche de classes d'usage des codons (contingence + AFC)

The screenshot displays the GenoBool software interface. An 'Import' dialog box is open, showing a list of variables for selection. The main window displays a table titled 'Query CDS_gbk' with the following data:

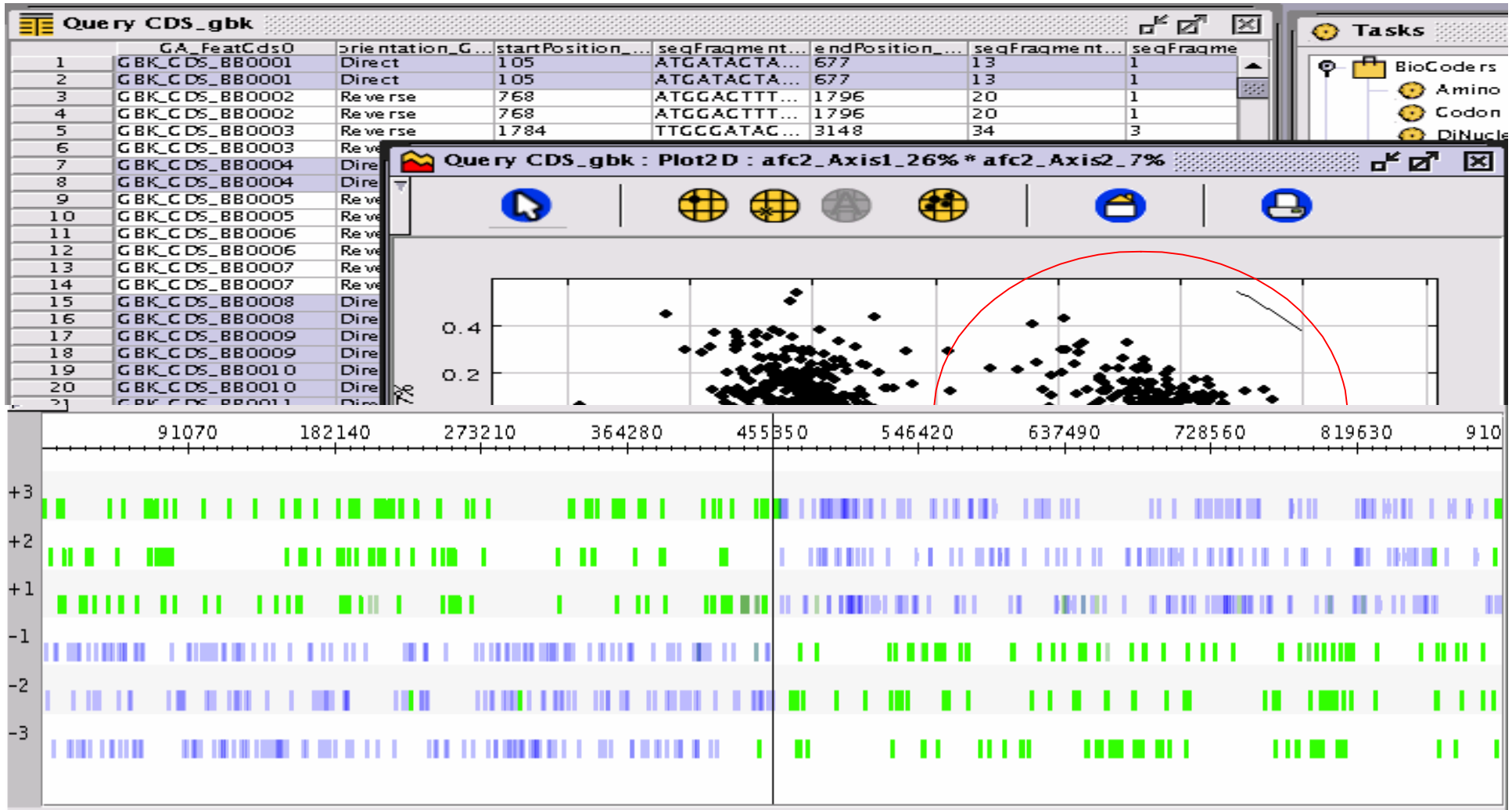
	on_G...	startPositio...	seqFragment_GA_IsLoc...	endPosition_...	seqFragment_GA_IsLocatedOn12:AAA_K	seqFi
1		105	ATGATACTAATGAAGT...	677	13	1
2		105	ATGATACTAATGAAGT...	677	13	1
3		768	ATGGACTTTTTAAAAA...	1796	20	1
4		768	ATGGACTTTTTAAAAA...	1796	20	1
5		1784	TTGGGATACCATGAA...	3148	34	3
6		1784	TTGGGATACCATGAA...	3148	34	3
7		3395	ATGCTTAAACAATATT...	5188	60	19
8		3395	ATGCTTAAACAATATT...	5188	60	19
9		5251	TTGATTTGAAAAGAA...	6312	25	2
10		5251	TTGATTTGAAAAGAA...	6312	25	2
11		6309	TTGATAAAAGCAGTA...	7433	18	2
12		6309	TTGATAAAAGCAGTA...	7433	18	2
13		7458	ATGAATGAAAGAGAA...	8315	14	1
14		7458	ATGAATGAAAGAGAA...	8315	14	1
15		8412	TTGAGTAGAAAATTTA...	9197	20	3
16		8412	TTGAGTAGAAAATTTA...	9197	20	3
17		9202	ATGGAATAGGCAAAA...	10206	43	11
18		9202	ATGGAATAGGCAAAA...	10206	43	11
19		10203	ATGAAGTCAATAGGA...	10577	16	0
20		10203	ATGAAGTCAATAGGA...	10577	16	0
21		10581	ATGAAAAAATATGA...	11420	39	10
22		10581	ATGAAAAAATATGA...	11420	39	10

The 'Tasks' panel on the right shows a tree structure with 'BioCodeurs' and 'Analysis' categories. 'Codon usage' is highlighted with a circled '1', and 'AFC' is highlighted with a circled '2'. Below the tasks panel, the 'Codon usage' section includes a link to 'documentation' and the text: 'Compute codon usage from a nucleotidic sequence'.

Interoperabilité GenoBool->GenoAnnot (2)



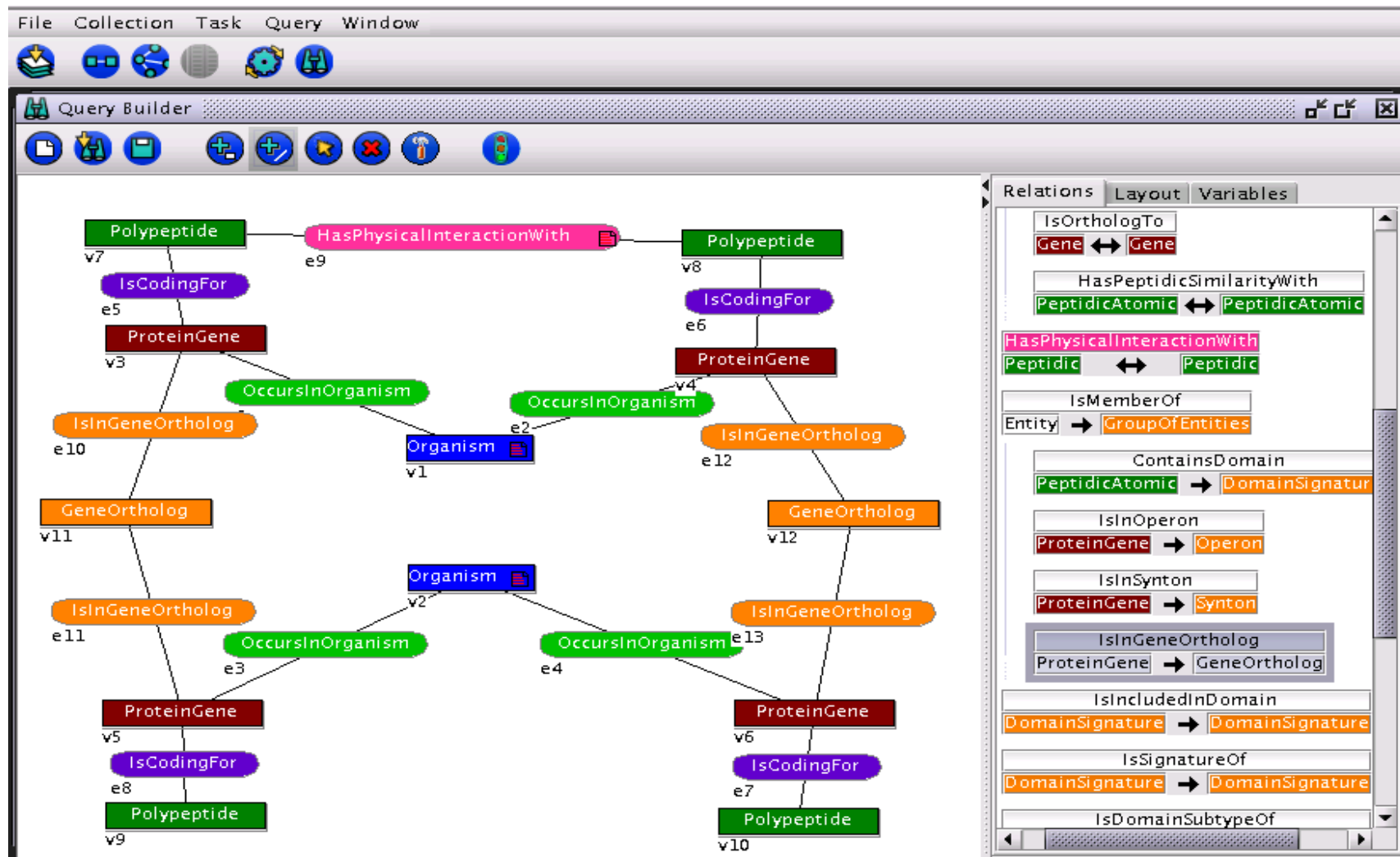
- Identification des sélectionnées CDS dans GenoAnnot
- Visualiser les CDS leading vs lagging



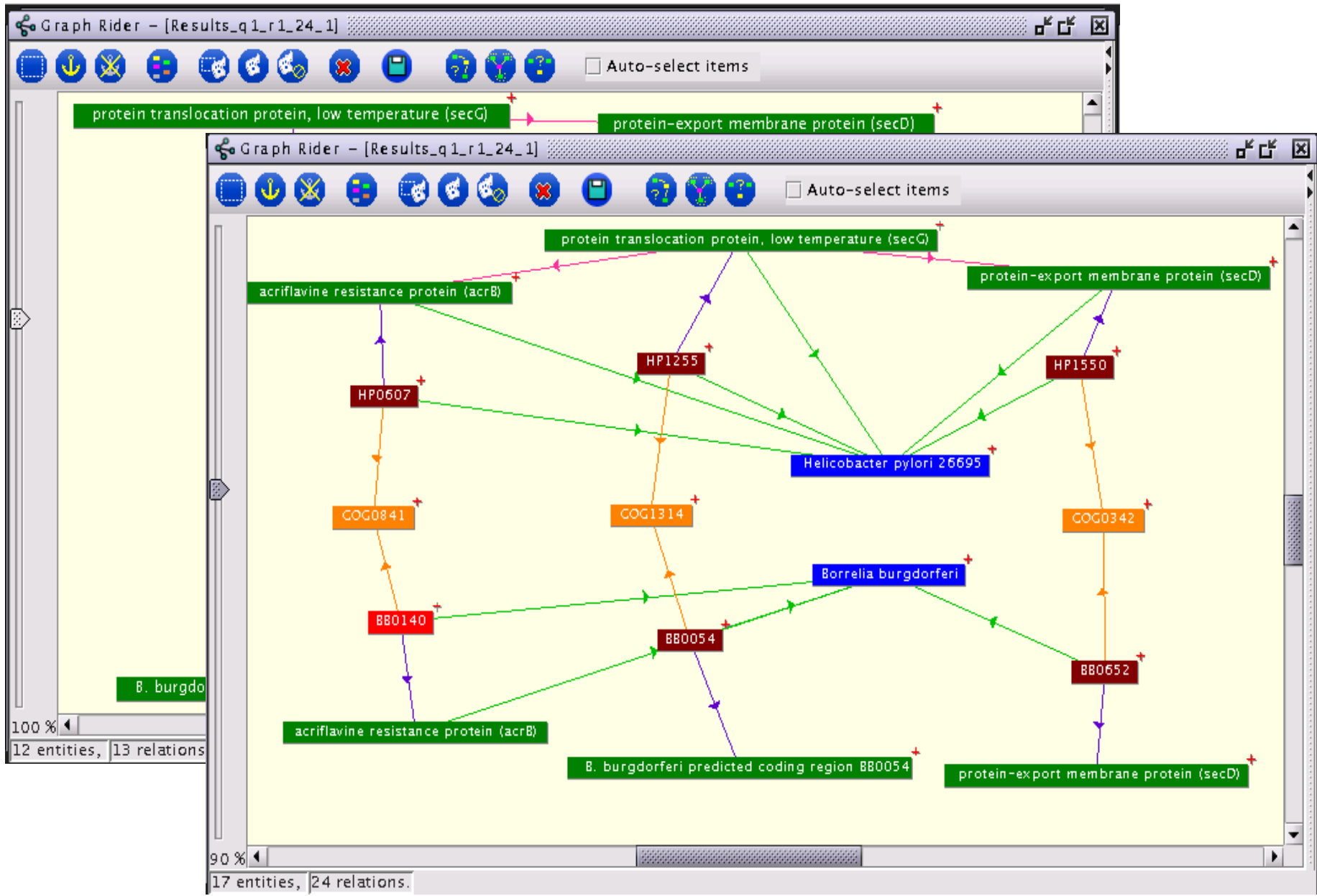
GenoLink : exploration du voisinage (1)



Espace de données : gènes, polypeptides, COGs, PIM pylori (Hgx)



GenoLink : exploration des voisinages (2)





- Genostar est mis à disposition des laboratoires de recherche publics

Décembre 2002 : Version 1.1

Juin 2003 : Version 1.2 (interopérabilité, ajout tâches...)

Décembre 2003 : Version 2.0

- programmable : ajout de méthodes, création de stratégie...

Actuellement en alpha-test puis beta-test

Pour en savoir plus...



<http://www.genostar.org>

Demande : academic@genostar.org , industry@genostar.org

Support / Retours utilisateurs : report@genostar.org



Le consortium Genostar





Christophe Bruley, INRIA, puis Genome express

Pierre-Emmanuel Ciron, INRIA

Antoine Danchin, Institut Pasteur

Stéphane Declere, Genome express

Jean-Louis Divol, Hybrigenics

Véronique Dupierris, INRIA

Patrick Durand, Hybrigenics

Gilles Faucherand, Genome express

Agnès Iltis, INRIA

Laurent Labarre, Hybrigenics

Claudine Medigue, Institut Pasteur

Alain Meil, Hybrigenics

Anne Morgat, INRIA

François Rechenmann, INRIA

Hélène Rivière-Rolland, Genome express

Vincent Schächter, Hybrigenics

Thierry Vermat, Genome express

Yves Vandembrouck, Genome express

Alain Viari, INRIA

Jérôme Wojick, Hybrigenics

