Pipeline d'annotation des variants génétiques

Sabrina Rodriguez
Maria Bernard

SIGEN@E

Equipe SIGENAE
Unité GABI, INRA de Jouy en Josas

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

16/01/2013

Plan

- ✓ **Contexte**
- ✓ **Quelques définitions**
- ✓ **Effets des SNPs sur les gènes**
- ✓ **Exemple de projet**
- ✓ **Description du pipeline d’annotation fonctionnelle**
- ✓ **Le format des résultats**
- ✓ **Conclusion et perspectives**

Contexte

✓ Dans le département de génétique animale, certaines équipes recherchent la cause génétique de phénotypes particuliers (maladies génétiques, corne chez les vaches).

✓ Avec l'avènement des techniques de séquençage haut débit, les génomes d'animaux d'intérêt sont séquencés et les variants génétiques (SNPs, insertions / délétions) prédits.

Certains des variants obtenus sont responsables de caractères recherchés.

Quelques définitions

Variation génétique / variant / polymorphisme:

un changement de base(s) à une même position sur le génome

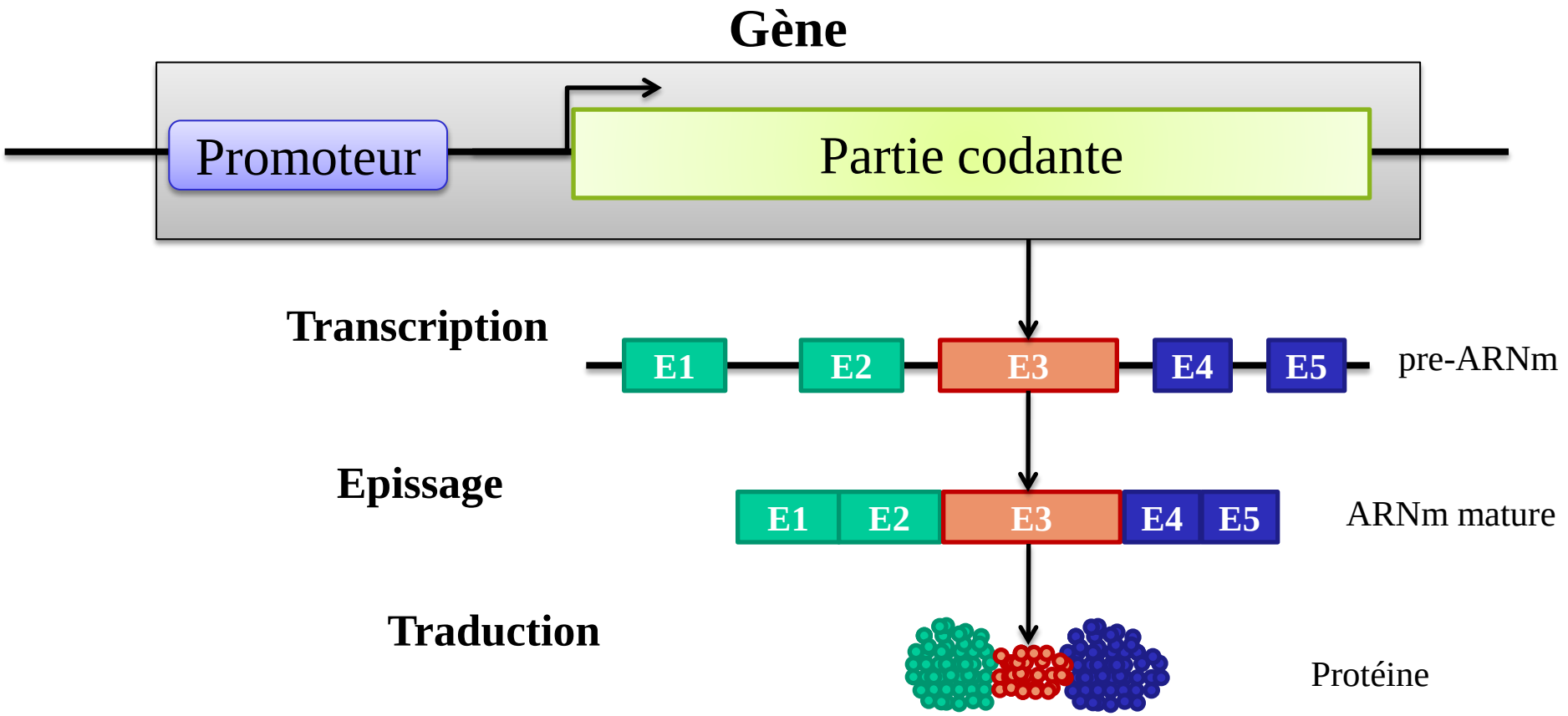
SNP (« *Single Nucleotide Polymorphism* ») : substitution d'un nucléotide dans une séquence nucléique.

rs42985251 (chromosome 1, position 47 828):

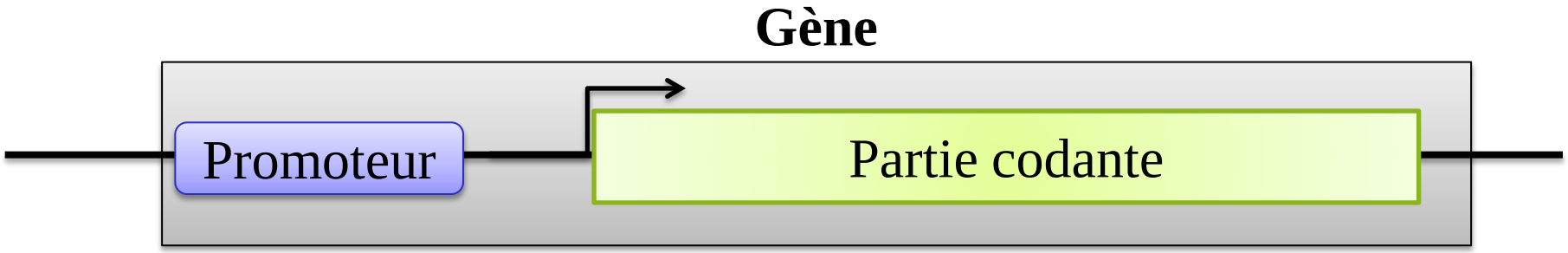
…AGGAA**G**CTGAC…

…AGGAA**A**CTGAC…

Quelques définitions



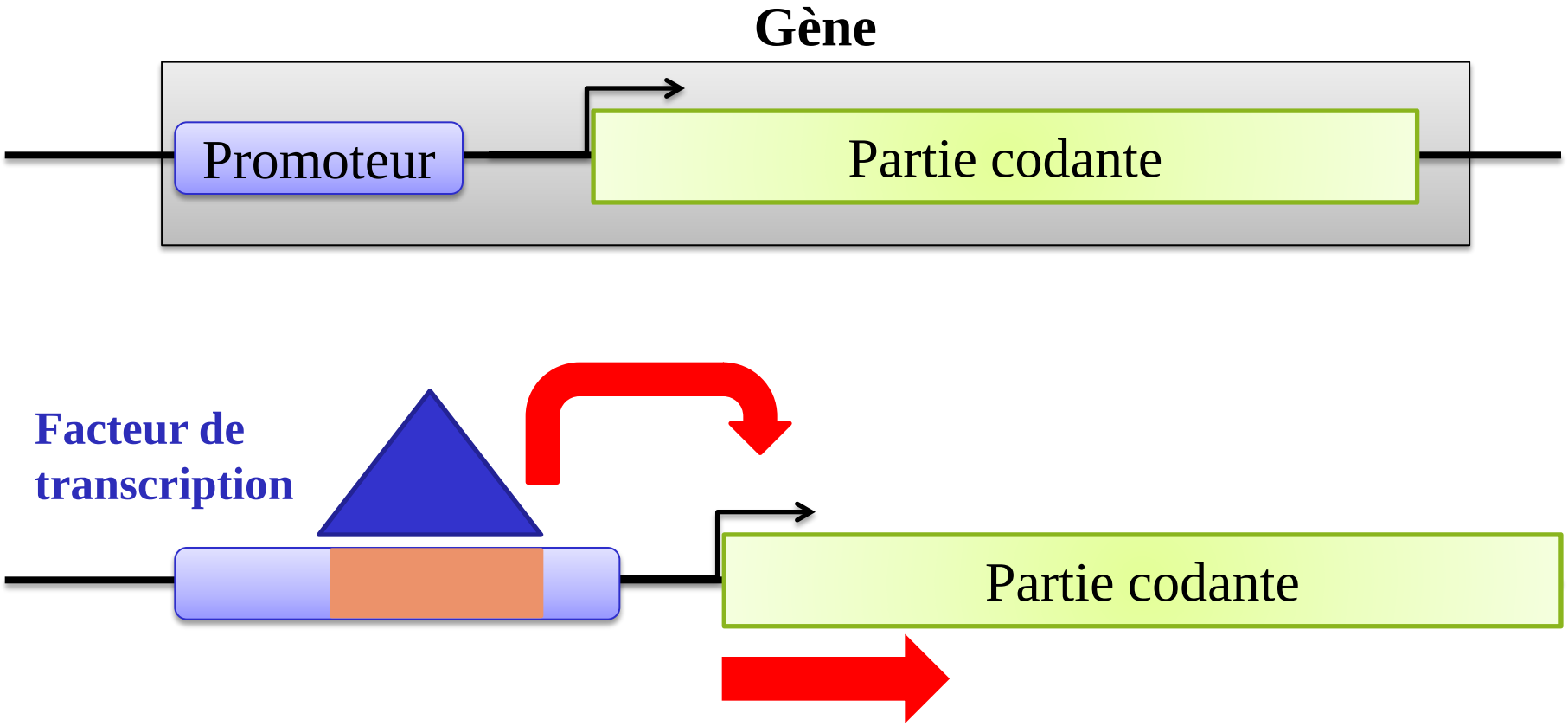
Effets des SNPs sur les gènes



Un SNP peut être localisé n'importe où sur le génome et affecter différents processus de biologie moléculaire

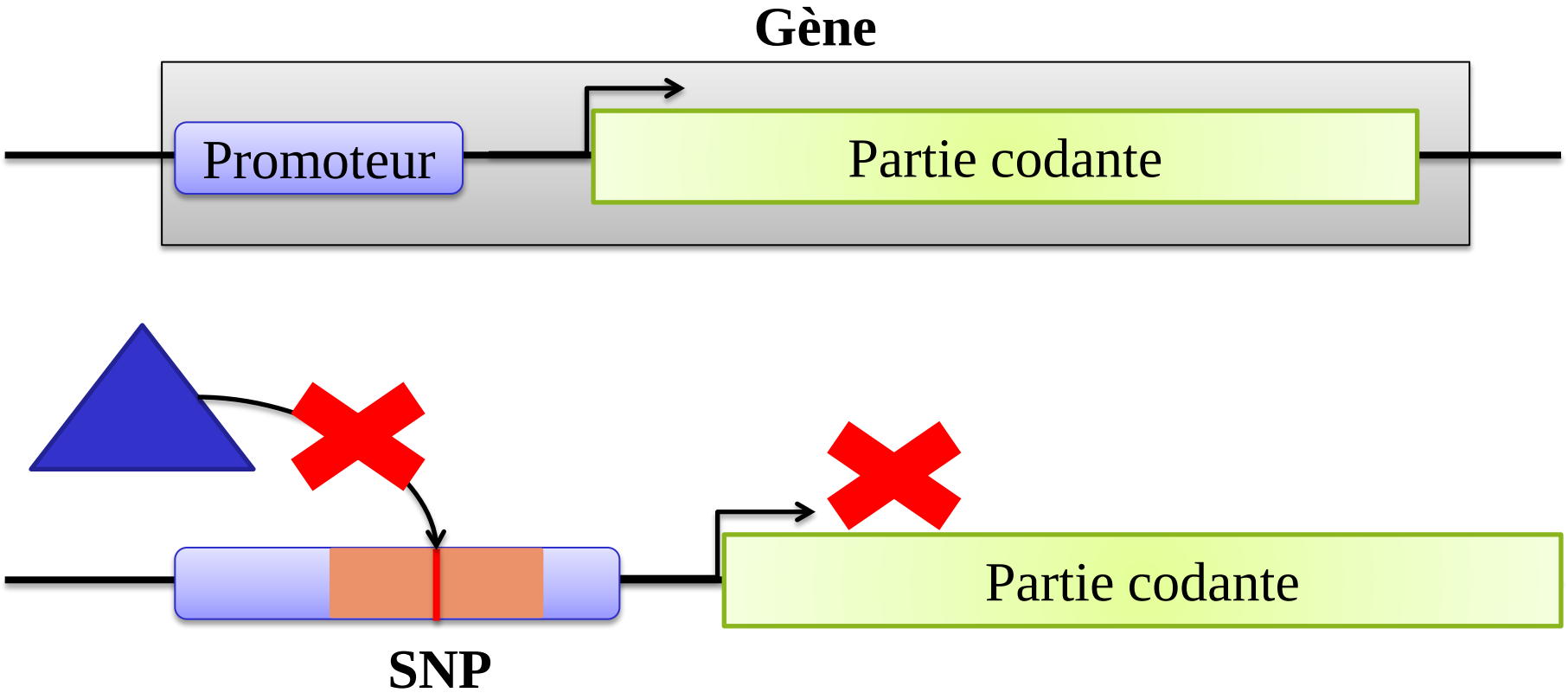
Effets des SNPs sur les gènes

Exemple de régulation de l'expression des gènes au niveau du promoteur :



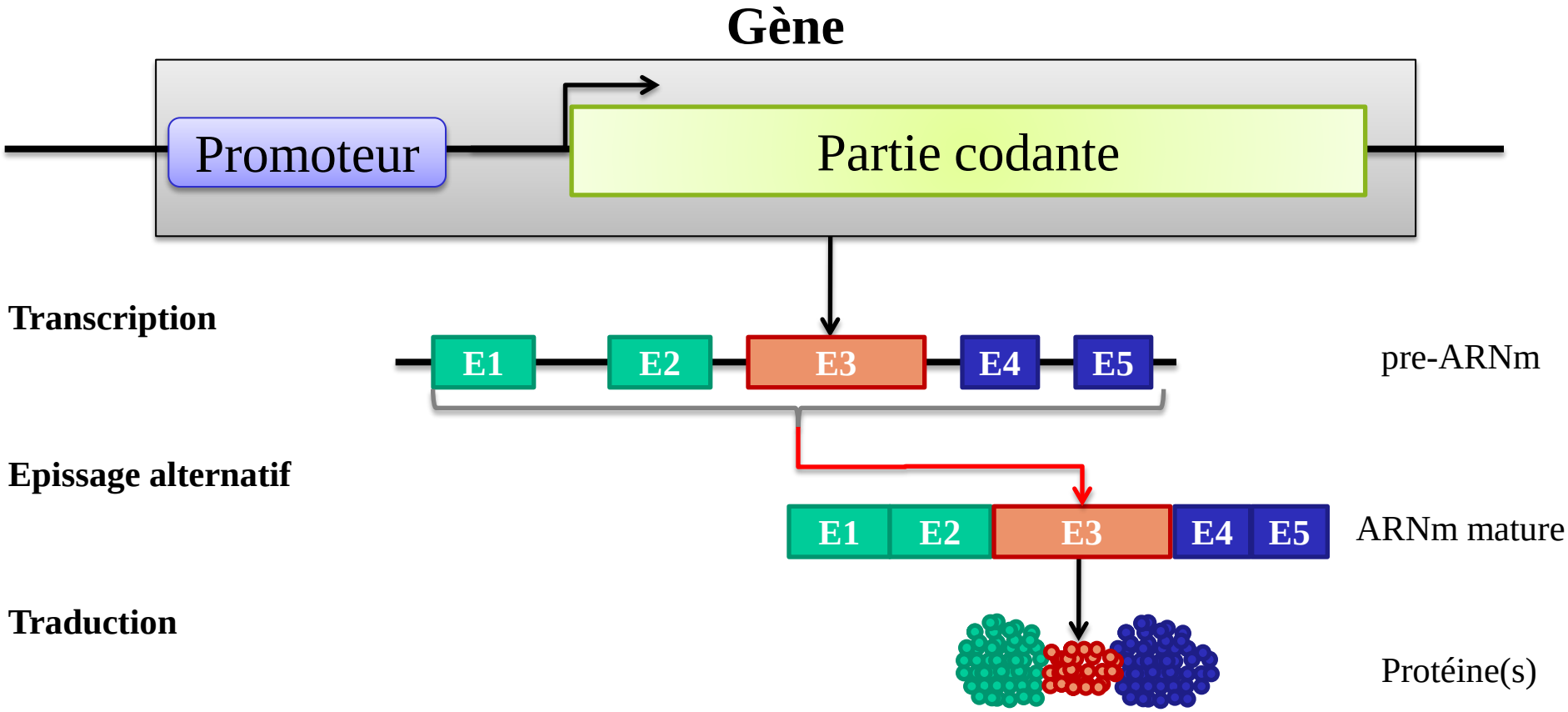
Effets des SNPs sur les gènes

Exemple de régulation de l'expression des gènes au niveau du promoteur :



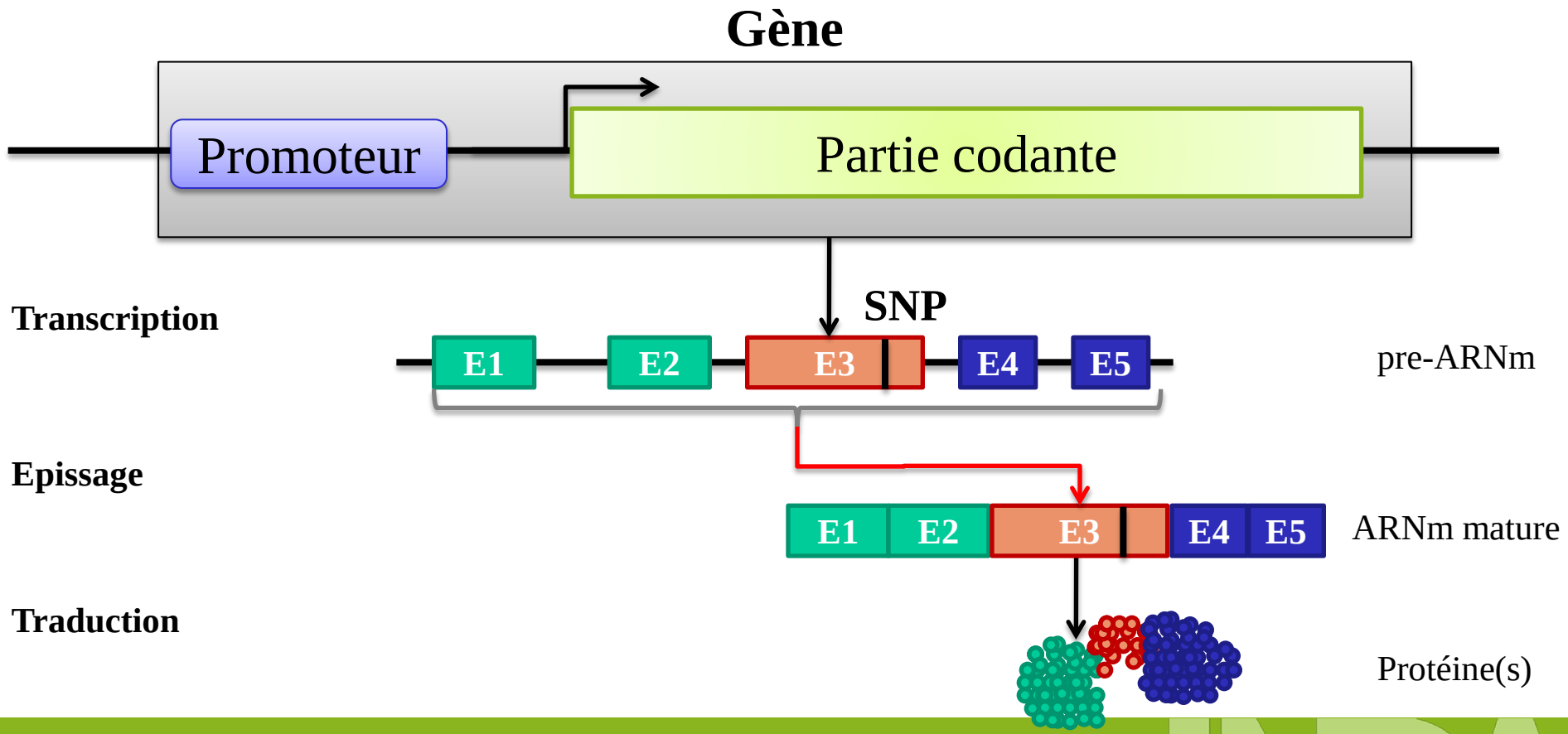
Effets des SNPs sur les gènes

Exemple d'effet sur la transcription des gènes:



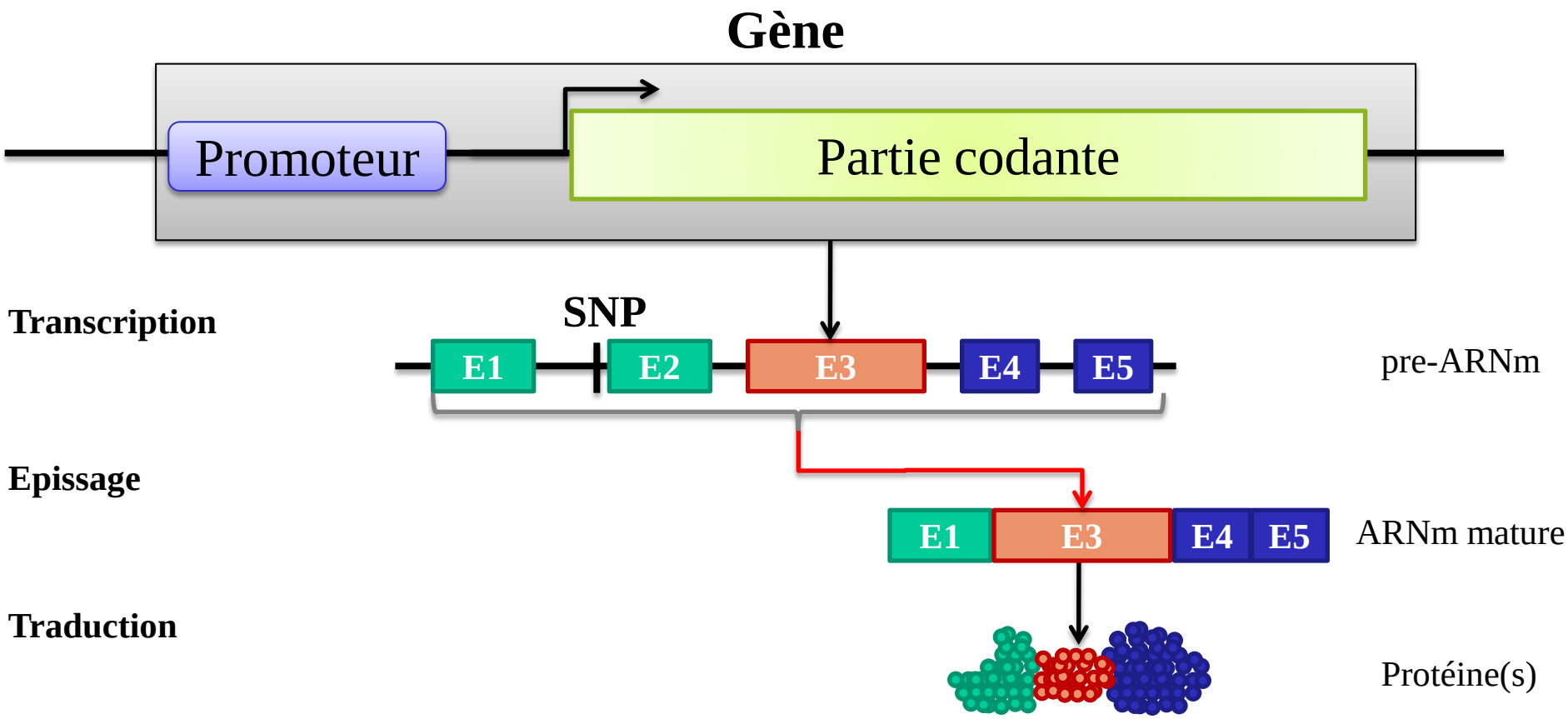
Effets des SNPs sur les gènes

Exemple d'effet sur la transcription des gènes:



Effets des SNPs sur les gènes

Exemple d'effet sur la transcription des gènes:



Exemple de projet

Maladie « Complex Vertebral Malformation » ou CVM

• Caractéristique : Veau avec ses vertèbres soudées

• Cause génétique : Anomalie génétique autosomique récessive dans l'espèce bovine , mutation dans le gène SLC35A3:

Substitution d'une base G en T

=> modification d'un acide aminé valine (V) en phénylalanine (F)

Thomsen B et al, Genome Res., 2006

```
ADN ch3  781 GAACTTTCAGCTGGCTCACAATTTGTAGGTCTCATGGCAGTTCTCACA
Prot     169 -E--L--S--A--G--S--Q--F--V--G--L--M--A--V--L--T-
                                                   TTT
                                                    F
```

=> « non synonymous coding SNP » ou « nscSNP »

Effets des SNPs sur les gènes

Modifications post-traductionnelles :

une **modification chimique d'une protéine** est réalisée le plus souvent par une enzyme, après sa synthèse ou au cours de sa vie dans la cellule.

Généralement cette modification entraîne un **changement de la fonction de la protéine** considérée, que ce soit au niveau de son action, de sa demi-vie, ou de sa localisation cellulaire.

Exemple:

- la **glycosylation** est **l'addition de glucides** aux chaînes peptidiques en croissance
  - La **O-glycosylation** est l'addition de glucides au niveau des résidus -OH des acides aminés **sérine** et **thréonine.**

Exemple de projet

Protocol général d'un projet d'analyse d'une anomalie génétique:

A partir de :

4 animaux séquencés (2 sains, 2 malades)

But:

déterminer le(s) polymorphismes responsables de la maladie

Exemple de projet

High Seq 2000

Ref ACGTACGTACGGCGTC

ACGTACGTACGGCGTC
ACGTACGTACGGCGTC
GTACGTACAGCGTC
CGTACAGCGTC

Technique utilisée:

Séquençage de génomes complets avec génome de référence (Plateforme GenoToul)

↓

Alignement des séquences sur génome de référence.

↓

Détection des variants potentiels (chez le bovin: >1 000 000 SNPs)

⇒ Comment sélectionner les SNPs potentiellement responsables des phénotypes recherchés et à valider en labo.?

⇒ Autrement dit, comment interpréter l'effet des SNPs sur l'organisme?

http://nextgenlab.cbi.buffalo.edu/?page_id=9
http://www.clcbio.com/index.php?id=785

Description du pipeline d'annotation fonctionnelle

- **Objectif:**
  - **A partir d'un ensemble de SNPs déterminer leurs annotations fonctionnelles** ou comment adapter les suites logiciels Pupasuite ou SNPnexus à d'autres espèces que l'homme et la souris

- **1ère étape :**
  - Analyse SNPs codants non synonymes, *ncsSNP*

Description du pipeline d'annotation fonctionnelle

- **Les ressources disponibles:**
  - des bases de données répertoriant les SNPs comme

    dbSNP : http://www.ncbi.nlm.nih.gov/projects/SNP/

    DGVa : http://www.ebi.ac.uk/dgva/
  - une multitude de programmes dédiés à des types très particuliers de modifications post-traductionnelles.
  - les API d'Ensembl

Description du pipeline d'annotation fonctionnelle

Pourquoi *Ensembl ?*

Ensembl est un système bio-informatique **d'annotation automatique de génomes**. C'est un projet conjoint de l'European Bioinformatics Institute (EBI) et du Wellcome Trust Sanger Institute dont l'idée centrale est **d'organiser de vastes champs d'information biologique autour de séquences génomiques**.

Ensembl se présente d'abord comme un navigateur de génomes (**Genome Browser**) permettant **d'explorer et de visualiser à différents niveaux les génomes de nombreux organismes**.

Ensembl est aussi une **base de données ouverte** dans laquelle on peut librement venir puiser, soit directement, soit à travers une interface de programmation (API), soit par le système d'interrogation BioMart.

Enfin, *Ensembl* est une infrastructure logicielle ouverte qui permet de construire différents systèmes organisant des données liées aux séquences génomiques.

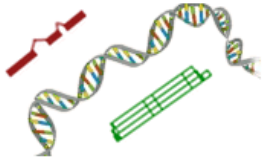
Description du pipeline d'annotation fonctionnelle

Ensembl les API

Les API ont pour but d'encapsuler la base de données en fournissant des accès aux différentes tables sans que leur schéma nécessite d'être connu par l'utilisateur.
Elles sont classés selon différentes thématiques.

Core

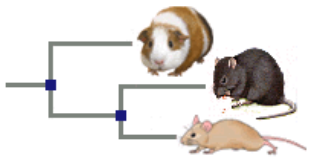
Sequence, genes and other automated annotation



- Perl API
- Database schema
- Tutorial

Comparative genomics

Homologues, paralogues and protein families



- Perl API
- Database schema
- Tutorial

Variation

SNPs, somatic mutations and structural variants

- Perl API
- Database schema
- Tutorial

Regulation

Regulatory features, motifs and oligoprobes

- Perl API
- Database schema
- Tutorial

Description du pipeline d'annotation fonctionnelle

Intérêts:

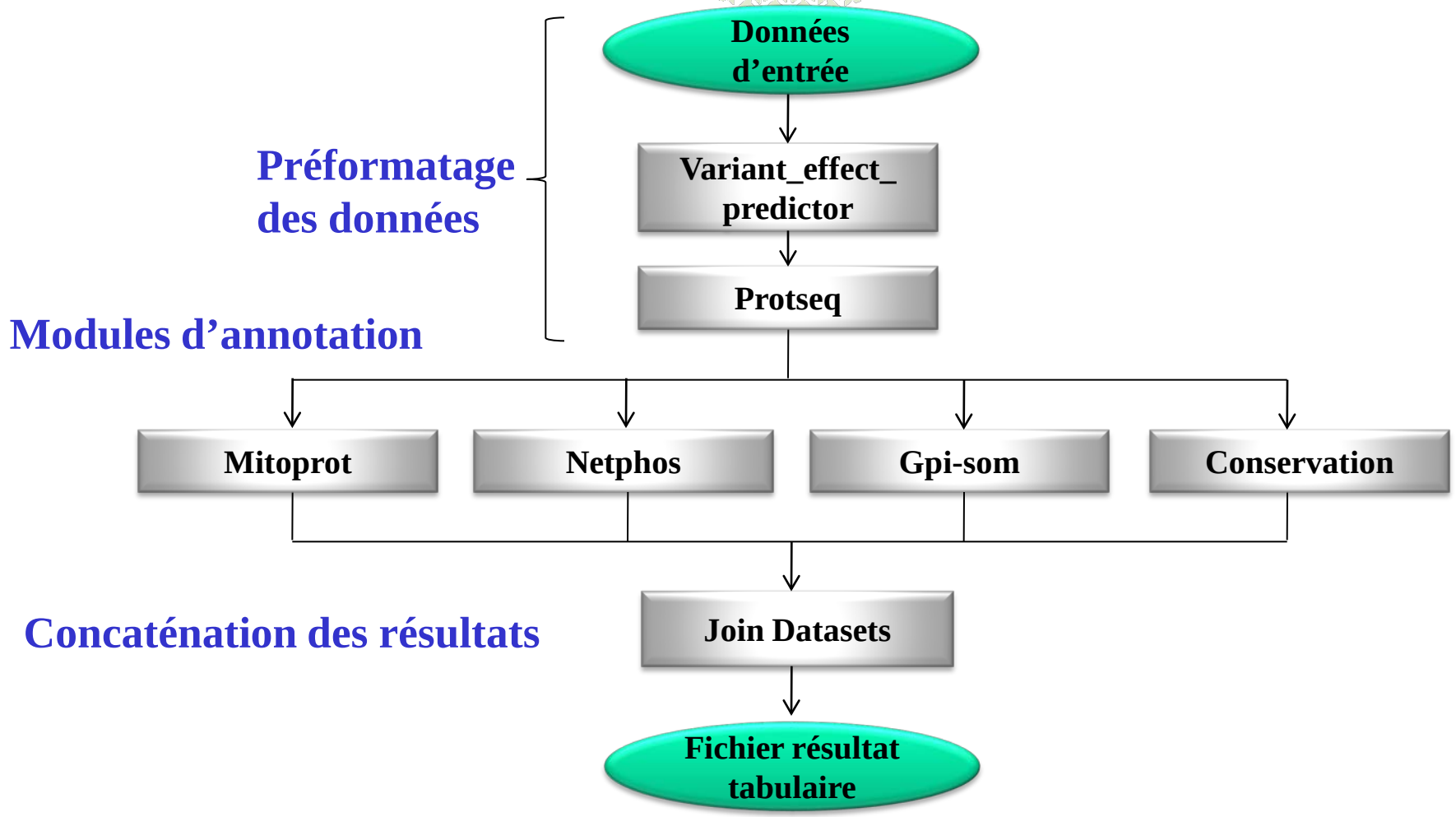
- Selon les espèces Ensembl fait appel à différentes bases de données, tant pour les SNP que pour les annotations.

Bos taurus

Source	Version	Description
dbSNP	133	Variants (including SNPs and indels) imported from dbSNP
DGVa	06/2012	Database of Genomic Variants Archive
Archive dbSNP	133	Former variants names imported from dbSNP

- Les BDs et APIs peuvent être installées en local pour une utilisation en fonction des besoins, une gestion des données de SNPs locale et un gain de temps d'interogation de la BD.

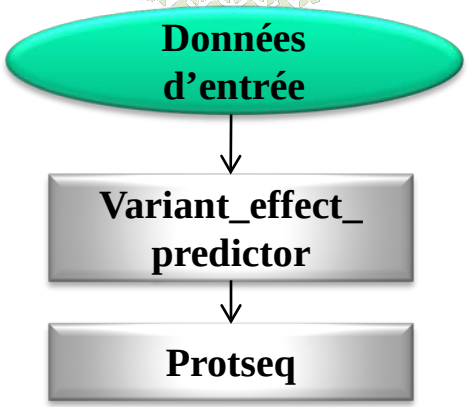
Description du pipeline d'annotation fonctionnelle



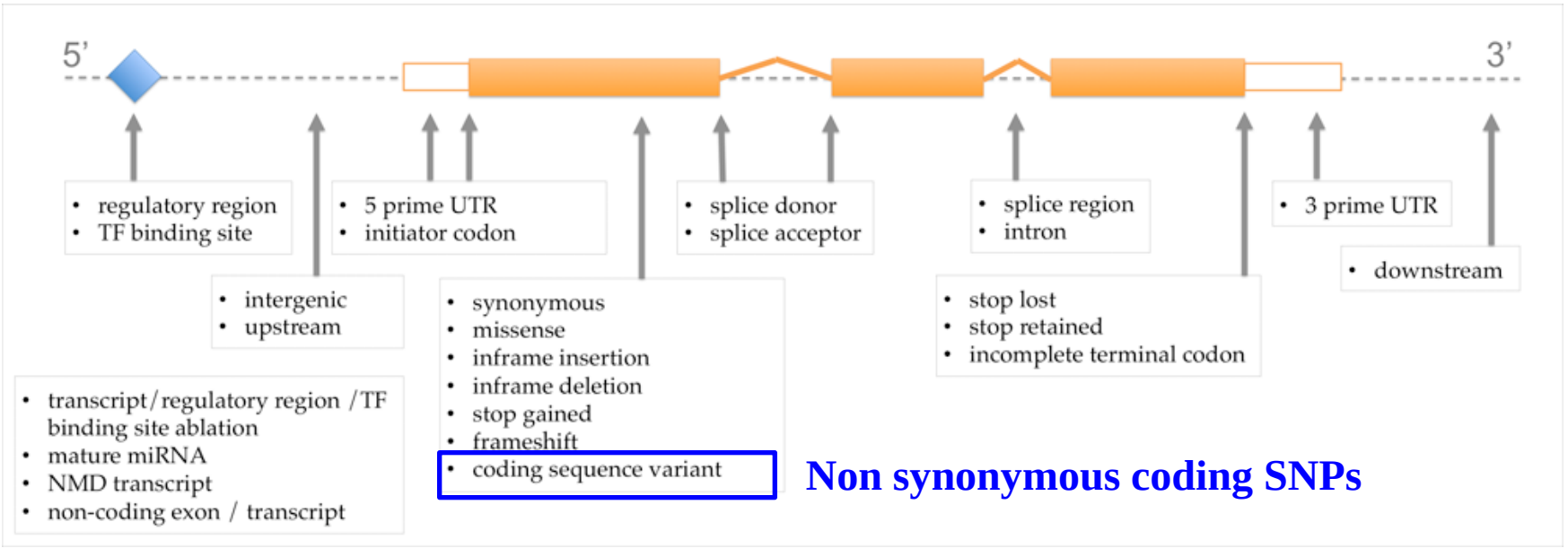
Source: personnelle

Description du pipeline d'annotation fonctionnelle

Préformatage des données



• **utilisation de l'outil « Variant_ effect_predictor »:**
Produit une annotation génomique des SNPs en fonction de leur localisation sur les gènes



Modules perl développés

Fichiers d'Input / d'output

http://www.ensembl.org/info/docs/variation/vep/index.html

Source: personnelle

Pipeline : 1) Préformatage des données

Format d'entrée de variant_effect_predictor:

- **Fichier VCF personnel**
- EnsEMBL défaut:

CHR	START	END	ALLELE	STRAND	IDENTIFIER*
5	140532	140532	T/C	+	
1	881907	881906	-/C	+	

*(optionnal) If not provided, the *runsnppredictor* will construct an identifier from the given coordinates and alleles.

- rsIDs (identifiants issus de la base de données publique dbSNP)
  ex: rs43580136

http://www.ensembl.org/info/docs/variation/vep/vep_formats.html

Pipeline : 1) Préformatage des données

Rappel du format VCF

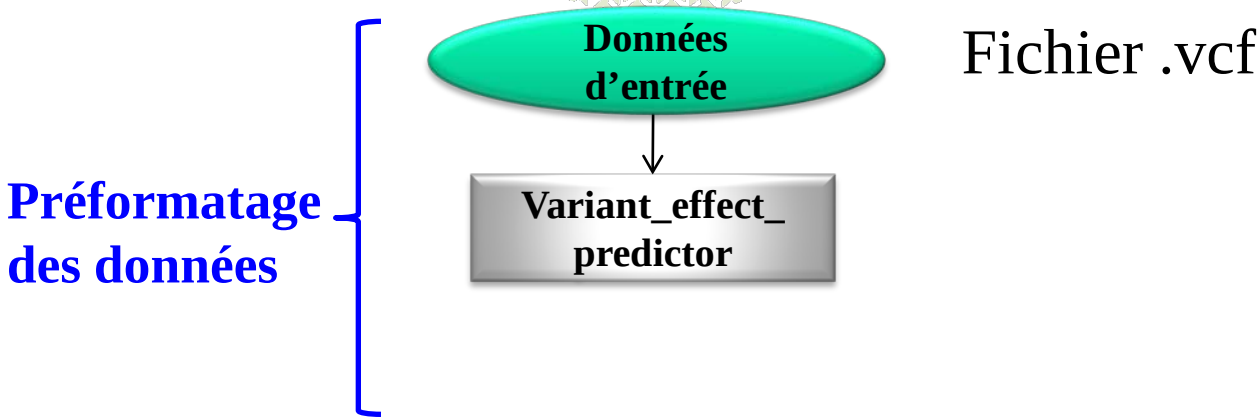
```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	ANIM1	ANIM2	ANIM3
20	**14370**	rs6054257	**G**	**A**	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	**1110696**	rs6040355	**A**	**G,T**	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4

http://www.1000genomes.org/node/101

Données d'entrée

Description du pipeline d'annotation fonctionnelle



Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Pipeline : 1) Préformatage des données

Ligne de commande du script « *runsnppredictor* »:

perl variant_effect_predictor.pl
--input_file variants.vcf | variants.ens | variants_rsID.txt
--output_file variants_consequences.vep
--registry ensembl.registry
--format vcf | ensembl | ID
--check_existing
--buffer_size 500
…

Pipeline : 1) Préformatage des données

Résultats de runsnppredictor

ENSEMBL VARIANT EFFECT PREDICTOR v2.6
Output produced at 2012-06-16 16:09:38
Connected to homo_sapiens_core_68_37 on ensembldb.ensembl.org
Using API version 68, DB version 68
Extra column keys:
DISTANCE : Shortest distance from variant to transcript

Colonnes 1 à 6
pour la descriptions du variants sur le génome

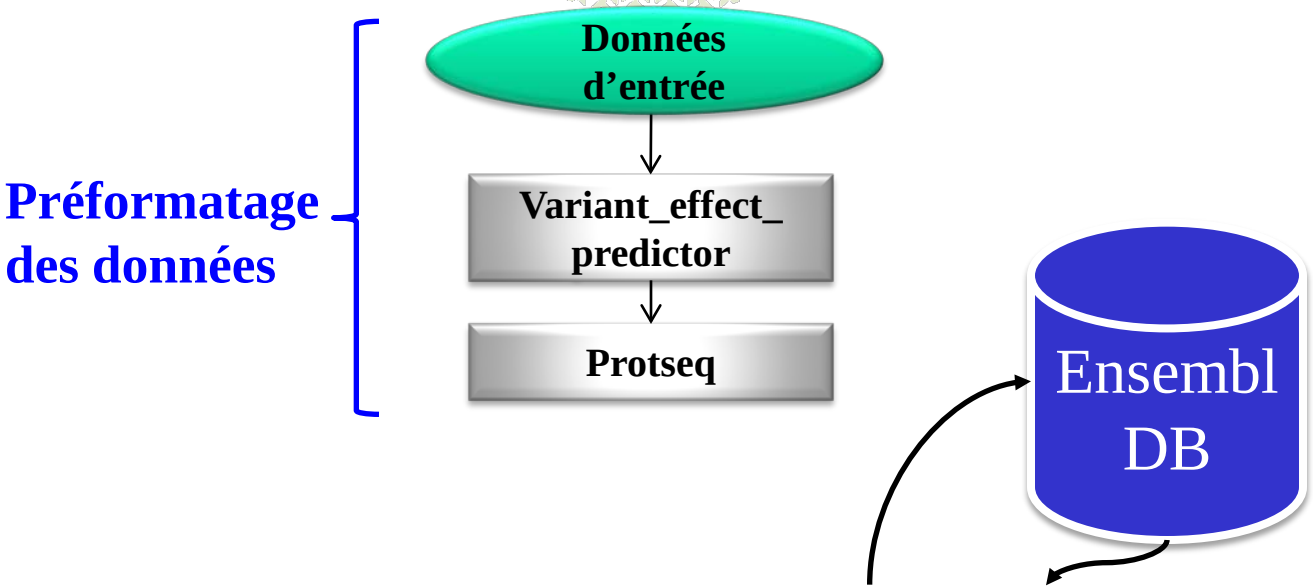
ID	pos	allele	gene	feature	feature type
11_224088_C/A	11:224088	A	ENSG00000142082	ENST00000525319	Transcript
11_224088_C/A	11:224088	A	ENSG00000142082	ENST00000534381	Transcript

Colonnes 7 à 14
Pour la description du variant sur le transcript et la protéine

Conséquences	Pos in cDNA	Pos in CDS	Pos in Prot	aa change	Codon change	Variation ID	Extra
NON_SYNONYMOUS_CODING	742	716	239	T/N	aCc/aAc	-	-
5_PRIME_UTR	-	-	-	-	-	-	-

http://www.ensembl.org/info/docs/variation/vep/vep_script.html#hgvs

Description du pipeline d'annotation fonctionnelle



Protseq : génère les séquences protéiques aux formats requis par les programmes

Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Pipeline: 1) Préformatage des données

Protseq

- **Objectif** : Récupérer les séquences protéiques impactées par les variants sélectionnés en utilisant les APIs d'Ensembl
- **Sortie** : Génère deux séquences protéiques par transcrit / SNP / allèle : une pour l'allèle de référence et 1 pour l'allèle alternatif considéré dans un fichier .fasta

Format de sortie :

> 7_14850002_G/T_rs43702456_ENSBTAT00000007344[D/E-12]_allele1 (REFERENCE)
MSGAIFTSLEGDGALDGTSGHPLVCPLCHAQYERPCLLDCFHEFCAGCLRGRAADGRLACPLCQHQTV
> 7_14850002_G/T_rs43702456_ENSBTAT00000007344[D/E-12]_allele2 (ALTERNATIF 1)
MSGAIFTSLEGEGALDGTSGHPLVCPLCHAQYERPCLLDCFHEFCAGCLRGRAADGRLACPLCQHQTV

> 7_14850002_G/T_rs43702456_ENSBTAT00000007344[D/E-12]_allele1 (REFERENCE)
MSGAIFTSLEGDGALDGTSGHPLVCPLCHAQYERPCLLDCFHEFCAGCLRGRAADGRLACPLCQHQTV
> 7_14850002_G/A_rs43702456_ENSBTAT00000007344[D/N-12]_allele2 (ALTERNATIF 2)
MSGAIFTSLEGNGALDGTSGHPLVCPLCHAQYERPCLLDCFHEFCAGCLRGRAADGRLACPLCQHQTV

Pipeline : 1) Préformatage des données

Ligne de commande de *Protseq :*

perl /BIN/formatSNPeffect-protseq.pl

--infile /DIR/variants_consequences.vep

--outfileprot /DIR/variants_consequences_proteins.fasta

--outfileprot_codedName /DIR/variants_consequences_proteins_encoded.fasta

--base_name "_seq"

--outencodingtable /DIR/names_encoding.txt

--species cow

--outvep_res /DIR/vep_results.txt

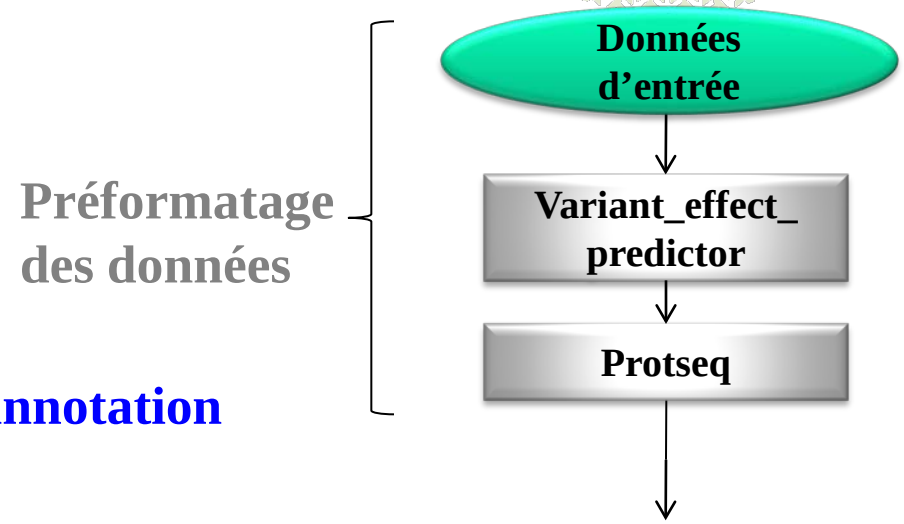
--outvep_title /DIR/vep_title.txt

--registry_file /usr/local/bioinfo/src/ensembl-api/variant_effect_predictor.registry

--out_vep_NSC_dir /DIR/ non_synonymous_coding.vep

…

Description du pipeline d'annotation fonctionnelle



Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Pipeline: 2) Annotation

Chaque module fait appel à un programme de prédiction (ex: de modification post-traductionnelle).

Pour chaque allèle on calcule un signal de prédiction.

Les annotations sont générées en comparant les signaux de l'allèle de référence (R) par rapport à l'allèle alternatif (A).

Annotation	Conditions			
	Signal référence	Signal alternatif	Score référence	Score alternative
Gain	NO	YES	$< \alpha$	$> \alpha$
Loss	YES	NO	$> \alpha$	$< \alpha$
Potential Loss	NO	NO	Score R – score A $> \beta$	
Potential Gain	NO	NO	Score A – score R $> \beta$	

Pipeline: 2) Annotation

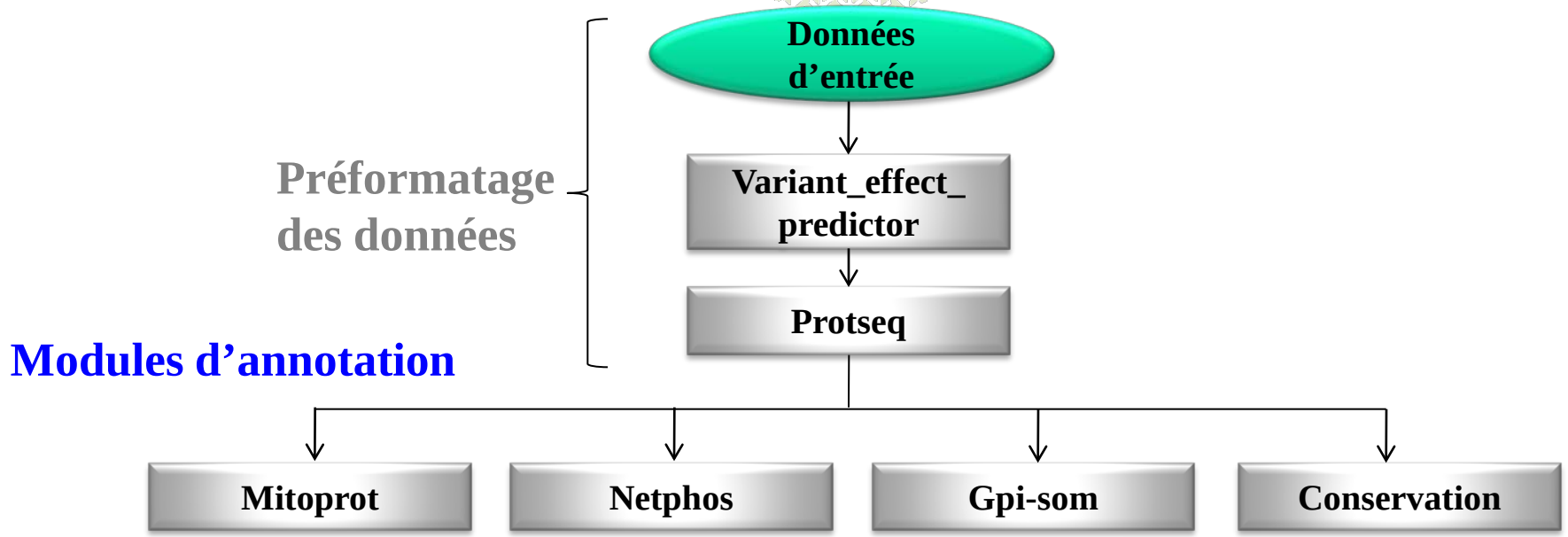
Exemple de recherche de différence de **phosphorylation** (module netphos) chez le bovin entre les deux variants

seuil paramétrable de détection de signal: $\alpha = 0.5$

seuil paramétrable de différence de score: $\beta = 0.3$

CHR	POS	Allèles	Score	Signal	CCL
1	100937768	C/T	0.2/0.639	no/W	Gain
1	103164202	A/T	0.55/no	W/no	Loss
1		C/T	0.1/0.4	no/no	Gain?
1		A/T	0.4/0.1	no/no	Loss?

Description du pipeline d'annotation fonctionnelle

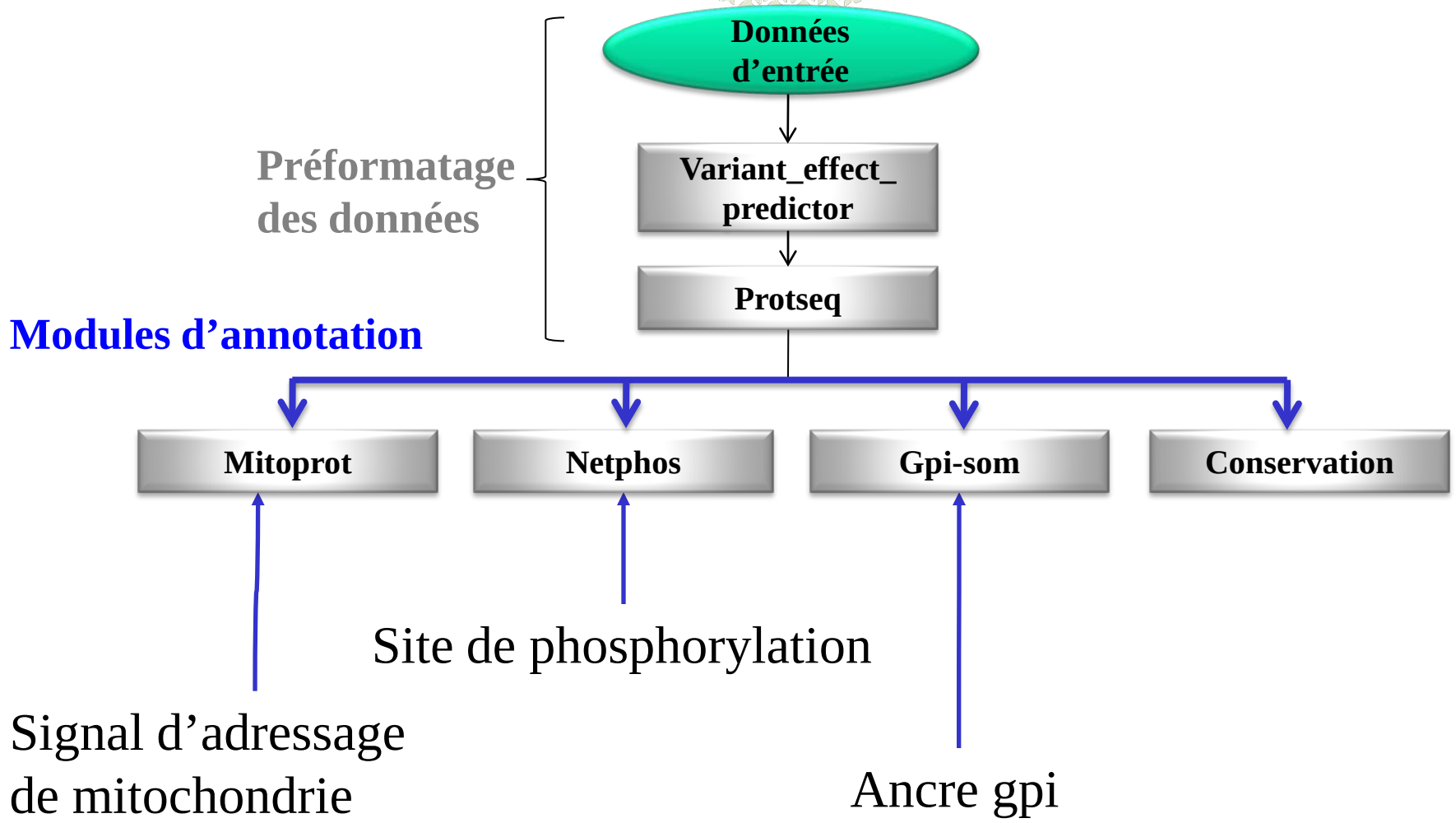


Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Description du pipeline d'annotation fonctionnelle

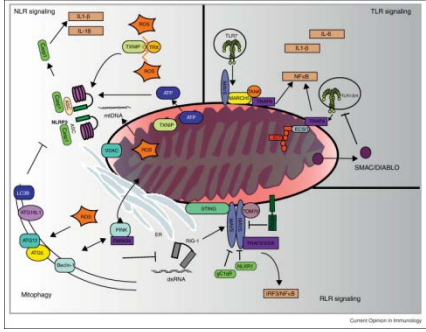


Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Pipeline: 2) Annotation Module Mitoprot

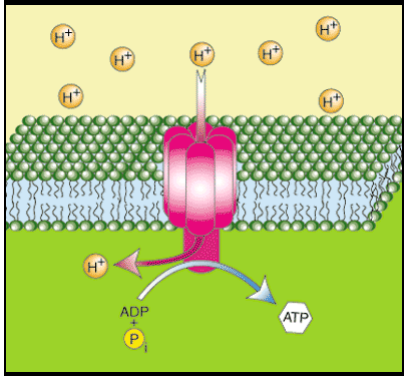


Description:
Ce composant vérifie s'il y a gain ou perte d'un signal d'adressage à la mitochondrie entre les séquences protéiques des 2 allèles par SNP et transcrit. Il utilise le programme « MITOPROT ».

Spécifications du programme:
- les séquences protéiques doivent commencer par une méthionine (M).
- une seule séquence est traitée à la fois.

Ligne de commande de *run_mitoprot.pl* :
perl run_mitoprot.pl
--infile /DIR/variants_consequences_proteins_encoded.fasta
--encodingfile /DIR/names_encoding.txt
--outpath mitoprot.out
--execommand /usr/local/bioinfo/src/mitoprot/mitoprotII-v1.101/mitoprot
--delta 0.5
--workpath /DIR

Pipeline: 2) Annotation Module Netphos



Description:

Netphos est un outil de prédiction des sites de phosphorylation.

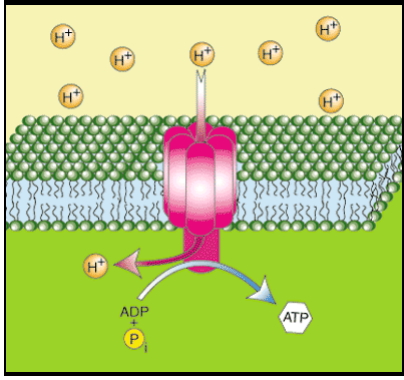
La phosphorylation consiste en l'ajout d'un groupement phosphate sur un acide aminé alcoolique (Serine, Thréonine et Tyrosine). Les enzymes responsables de la phosphorylation sont les kinases. Pour chaque acide aminé alcoolique, Netphos détermine un score de phosphorylation pour chaque enzyme (20 enzymes traitées au total). Les enzymes sont : ATM, CaMII, cdc, cdk, CKI, CKII, DNAPK, EGFR, GSK3, INSR3, 38MAPK, PKA, PKB, PKC, PKG, RSK, SRC. La plupart étant des kinases.

http://www.cbs.dtu.dk/services/NetPhos/

http://www.sparknotes.com/biology/cellrespiration/oxidativephosphorylation/section3

Sequence and structure based prediction of eukaryotic protein phosphorylation sites.
Blom, N., Gammeltoft, S., and Brunak, S. **Journal of Molecular Biology: 294(5): 13511362, 1999.**

Pipeline: 2) Annotation Modules Netphos



Spécifications:

- 1 seule séquence protéique a la fois
- les séquences protéiques ne doivent pas dépasser 4000 acides aminés

Lignes de commande de run_netphos.pl :

Perl run_netphos.pl
--infile /DIR/variants_consequences_proteins_encoded_4000.fasta
--encodingfile /DIR/names_encoding.txt
--execommand /usr/local/bioinfo/bin/netphos-3.1
--delta 0.5
--outfile netphos.out

http://www.cbs.dtu.dk/services/NetPhos/

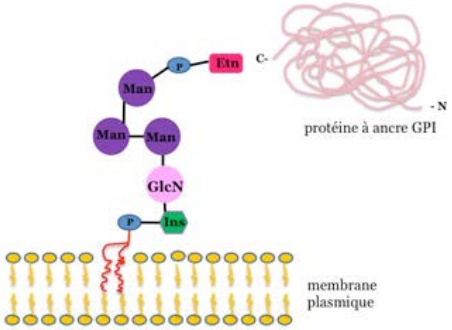
http://www.sparknotes.com/biology/cellrespiration/oxidativephosphorylation/section3

Sequence and structure based prediction of eukaryotic protein phosphorylation sites.
Blom, N., Gammeltoft, S., and Brunak, S. **Journal of Molecular Biology: 294(5): 13511362, 1999.**

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

Netphos

Pipeline: 2) Annotation Module GPI-SOM



Description:

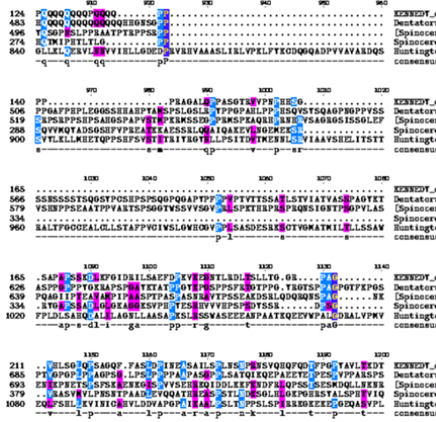
Ce composant vérifie s'il y a gain ou perte d'un signal d'ancre gpi (« GlycosylPhosphatidylInositol », ou glypiation) entre les séquences protéiques des 2 allèles par SNP et transcrit. Il utilise le programme « gpi-SOM ».

Ligne de commande :

Perl extract_gpi-som_results.pl
--infile /DIR/variants_consequences_proteins_encoded.fasta
--encodingfile /DIR/names_encoding.txt
--outpath gpi.out
--execommand /usr/local/bioinfo/src/kohgpi/current/kohgpi
--delta 0.5
--workpath /DIR

http://gpi.unibe.ch/
GPI-SOM: Identification of GPI-anchor signals by a Kohonen Self Organizing Map
http://www.micalis.fr/Poles-et-Equipes/Pole-Risques/vif_lavie_richard

Pipeline: 2) Annotation Module conservation



Description:

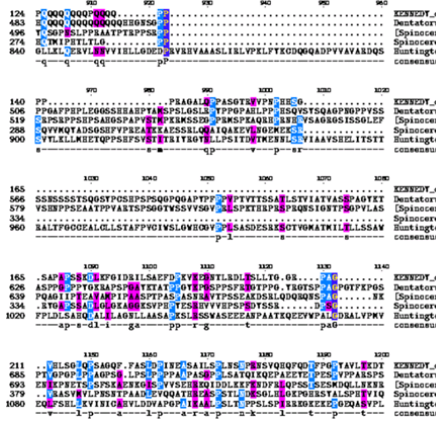
Ce module génère 2 résultats selon 2 méthodes différentes.

• La première méthode utilise l'API d'Ensembl Compara pour calculer la conservation d'un acide aminé à une position X d'une protéine par rapport aux protéines orthologues disponibles dans Ensembl

2 scores sont générés: un score observé et un score attendu.
Le score attendu prend en compte le contexte génomique pour donner une indication de conservation (région soumise à forte pression de sélection? Autrement dit hautement conservée dans les autres espèces).

Plus la différence de score est élevée, plus la position est conservée et plus une mutation à cette position a de fortes chances d'être délétère.

Pipeline: 2) Annotation Module conservation

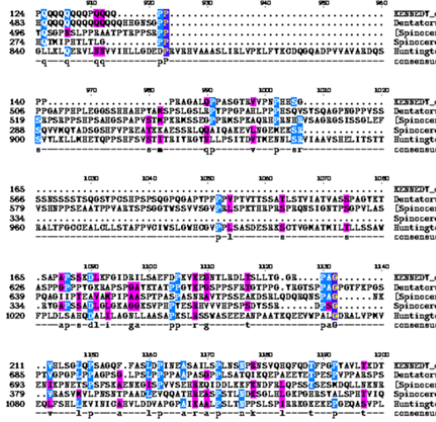


Description:

• La seconde méthode utilise la matrice Grantham. Cette matrice est une matrice de transition entre les 20 acides aminés connus en se basant sur les propriétés physico-chimiques et le volume moléculaire de chaque acide aminé. Plus le score est petit plus les propriétés physico-chimique de la protéine sont respectées:

- Score < 50; conservée
- Score entre 51 et 100; modérément conservée
- Score entre 101 et 150; modérément radicale
- Score >151 ; radicales

Pipeline: 2) Annotation Module conservation



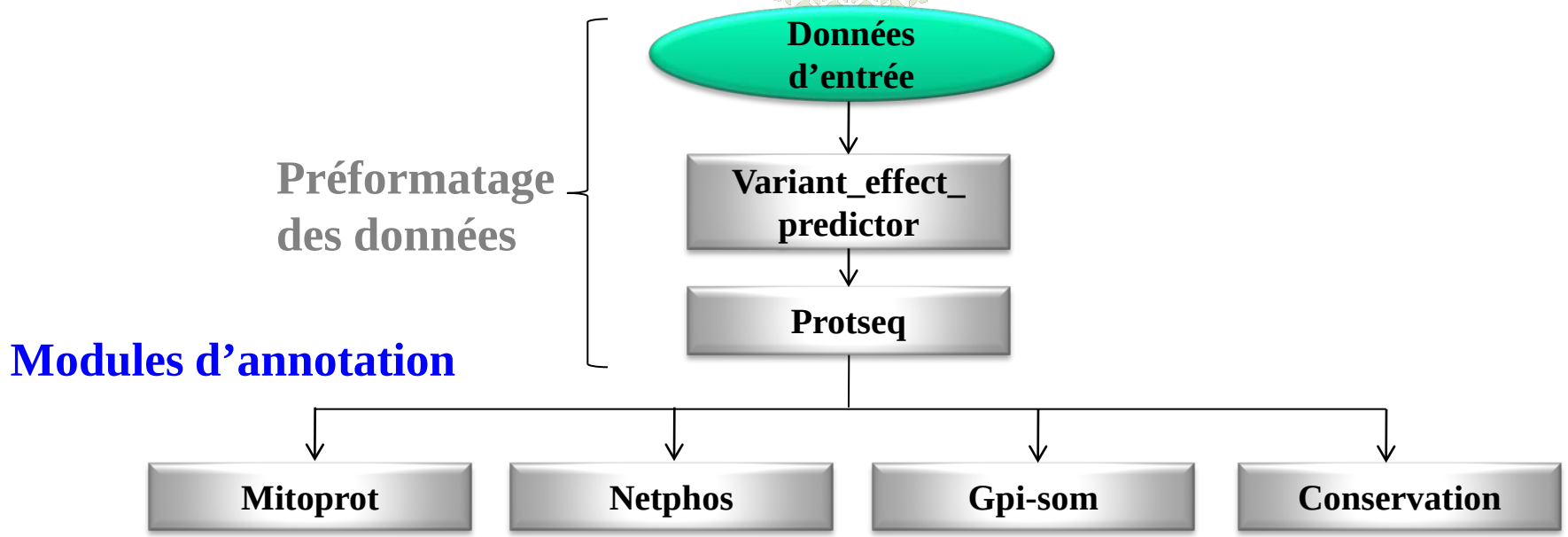
Spécifications du programme:

- utilise en entrée un fichier vep avec les SNPs « non_synonymous_coding »

Ligne de commande :

perl extract_conservation_results.pl
--infile /DIR/ non_synonymous_coding.vep
--compara_user compara67
--species cow
--encodingfile /DIR/names_encoding.txt
--workpath /DIR

Description du pipeline d'annotation fonctionnelle

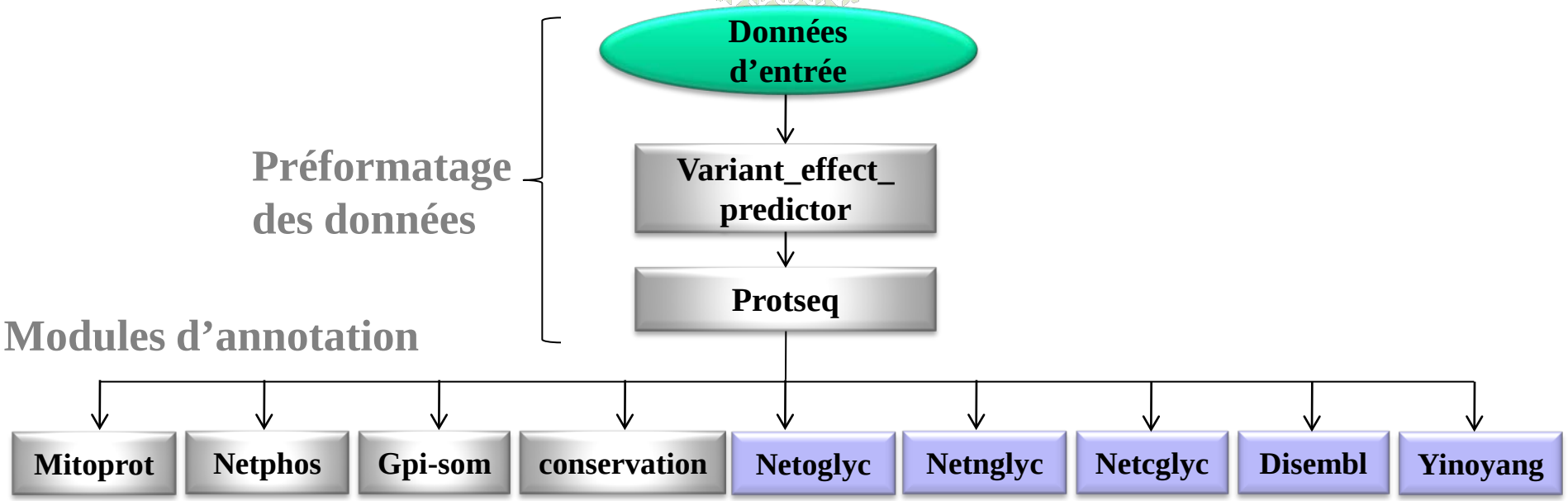


Modules perl développés

Fichiers d'Input / d'output

Source: personnelle

Description du pipeline d'annotation fonctionnelle



Modules perl développés

Fichiers d'Input / d'output

Modules perl développés (ergatis seulement)

Source: personnelle

Pipeline: 2) Annotation
Analyses additionnelles et spécifications

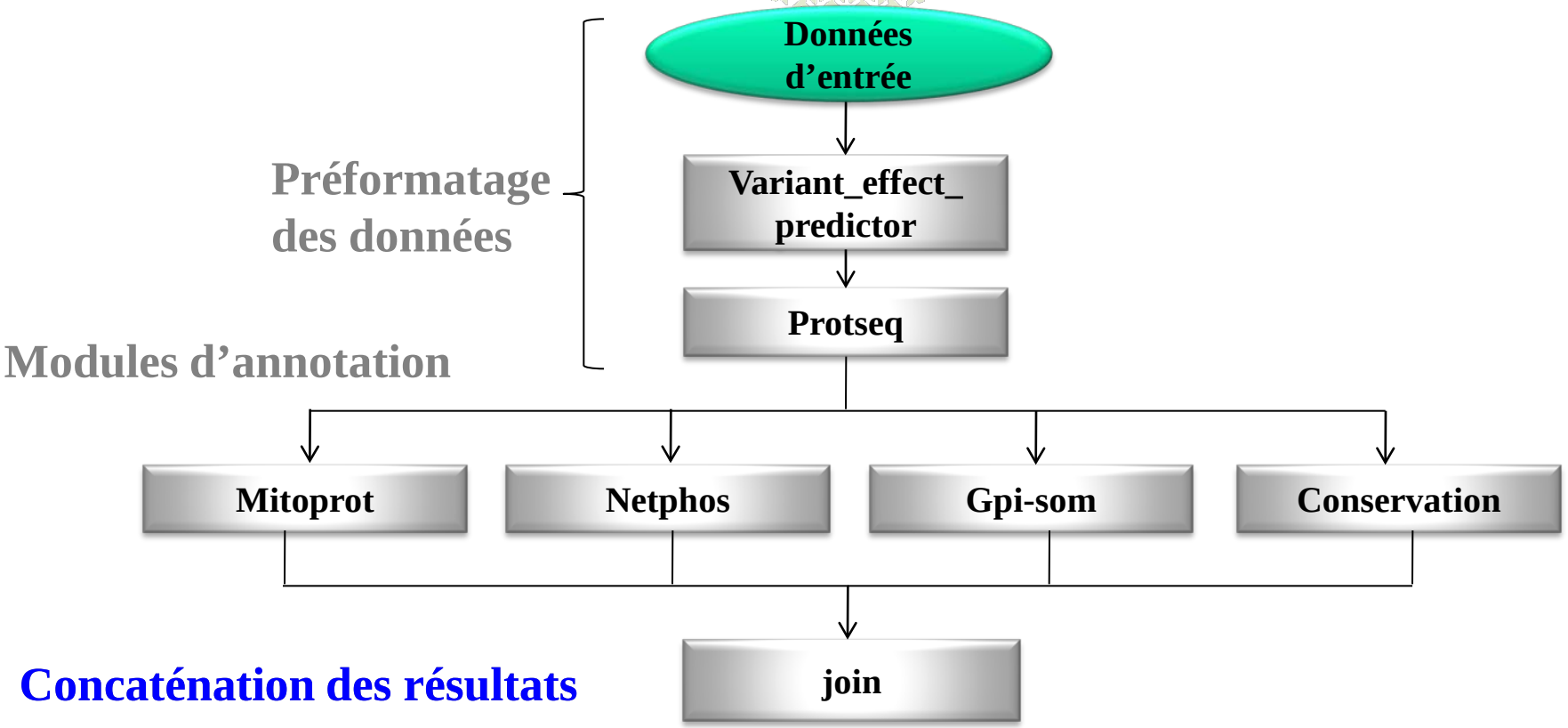
Les autres modules:

- **Netcglyc:** C glycosylation
- **Netnglyc:** N glycosylation
- **Netoglyc:** O glycosylation
- **Yinoyang:** predictions des site de fixation des O-ß-GlcNAc
- **Disembl:** désordre protéique

Certains modules sont restreints à un groupe de taxons

Gpi-som	mitoprot	netphos	netcglyc	netoglyc	netnglyc	yinoyang	disembl	conservation
all	eucaryote	eucaryote	mammifère	mammifère	humain	eucaryote	all	eucaryote

Description du pipeline d'annotation fonctionnelle

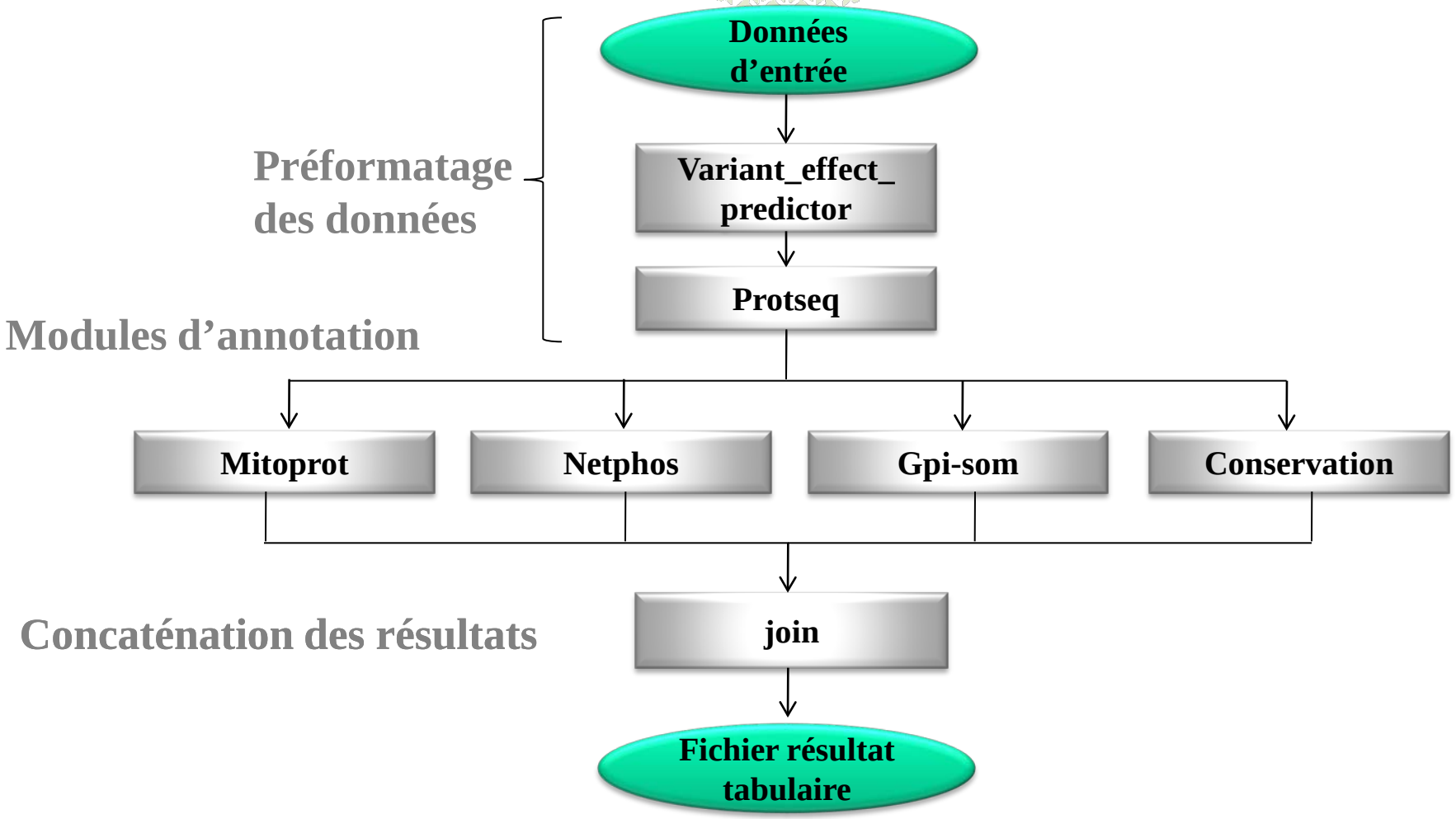


Modules perl développés

Fichiers d'Input / d'output

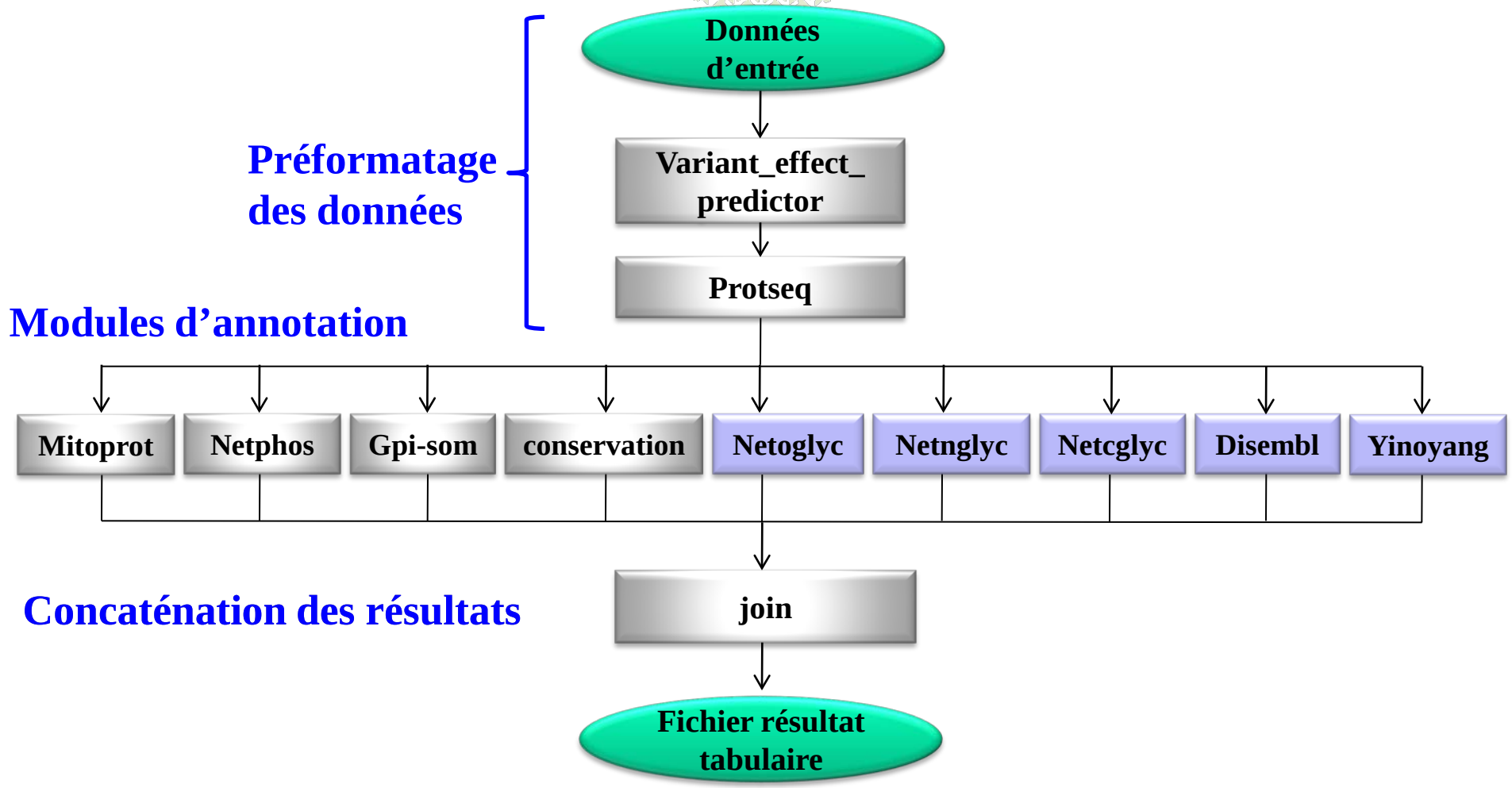
Source: personnelle

Description du pipeline d'annotation fonctionnelle



Source: personnelle

Description du pipeline d'annotation fonctionnelle



Source: personnelle

Pipeline: 3) Obtention des résultats

Join Datasets

- **Objectif** : Fusionner les résultats des différents modules pour chaque SNP
- **Sortie** : Un fichier tabulaire, regroupant toutes les conclusions des modules précédents

Name	conservation Grantham_score	conservation Observed_score	conservation Expected_score	conservation Difference_score	mitoprot Case	mitoprot score allele1	mitoprot score allele2	netphos ATM result	netphos ATM allele1 score
13_61584514_T/A_-_ENSBTAT00000027390[L/Q-11]	113	0.0000	0.9150	0.9150	gain	0.1882	0.2982		
1_154285564_C/A_-_ENSBTAT00000030369[S/I-29]	142	2.1600	3.9500	1.7900	loss	0.3758	0.4093		
1_154285564_C/A_-_ENSBTAT00000061345[S/I-29]	142	2.1600	3.9500	1.7900					
1_18073037_C/A_-_ENSBTAT00000000788[S/Y-167]	144	0.0000	3.7500	3.7500					
1_18073037_C/A_-_ENSBTAT00000064206[---]	0	0.0000	3.7500	3.7500					
1_42531330_C/T_-_ENSBTAT00000005398[E/K-29]	56				loss	0.0900	0.6499		
1_58593687_G/A_-_ENSBTAT00000057527[S/F-633]	155	0.8500	4.1900	3.3400				loss	0.553
1_58792664_G/A_-_ENSBTAT00000027966[S/F-1619]	155	0.0000	4.4400	4.4400				loss	0.529
1_68929240_C/A_-_ENSBTAT00000000740[W/L-80]	61	2.6800	4.1200	1.4400					
1_68929240_C/A_-_ENSBTAT00000054897[W/L-32]	61	2.6800	4.1200	1.4400	loss	0.3961	0.4230		
1_95205755_G/A_-_ENSBTAT00000052461[S/N-171]	46							loss	0.546

Conclusion

- Le pipeline de Galaxy inclut actuellement 4 programmes d'annotation (serveur local)

- le pipeline est aussi disponible sur galaxy SIGENAE et intégrera d'autres modules http://sigenae-workbench.toulouse.inra.fr/galaxy

- Annotation des nscSNPs bovins : sur un jeu de données de 1182 SNPs non synonymes codants, 28% ont un effet prédit

Perspectives

Développements à l'étude:

- ajout de nouveaux modules d'annotations fonctionnelles comme SIFT (analyse de la conservation)
- annotations de l'impact des SNP présents dans les sites de fixations aux TFBS
- annotations de l'impact des SNP présents dans les miRNAs
- annotations des SNPs responsables d'épissage alternatif

Merci de votre attention

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

BACKUP SLIDES

Développement du pipeline

Intégration de programmes de prédiction de modifications post traductionnelles développés par divers organismes:

- Glypiation ➔ GPI-SOM
- Adressage à la mitochondrie ➔ Mitoprot
- Phosphorylation ➔ Netphos (CBS server)
- Glycosylation ➔ NetNGlyc, YinOYang, NetCGlyc (CBS server)

Annotation fonctionnelle des SNPs localisés dans les régions régulatrices

- objectif:

L'annotation des SNPs en amont des gènes consisterait à prédire les sites de fixation des facteurs de transcription en amont des gènes à partir des motifs issus de différentes bases de données (Genomatix, Transfac, Jaspar…), et de comparer l'effet des des SNPs localisés au niveau de tels sites.

- Espèces concernées: bovin, cheval, autres espèces domestique

Définitions

- API : Application Programme Interface ou interface de programmation. Elle permet l'interaction des programmes les uns avec les autres (ici Perl et la base de donnée d’Ensembl). Du point de vue technique une API est un ensemble de fonctions, procédures ou classes.

Programmes

Les résultats :

-En général, avec le score :

Négatif < 0.5 < Positif

-Sinon :

Le programme l'indique par une note.

On indique une disparition/apparition d'un site si :

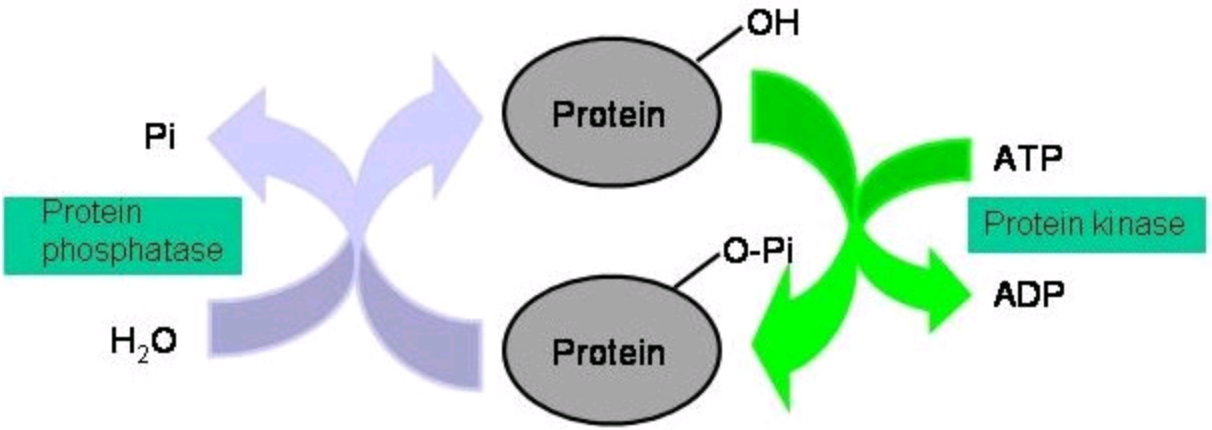
- La différence de score est trop importante entre les allèles ;
- Un des deux allèles n'est pas annoté par le programme.

Si score allèle 1 > score allèle 2 => disparition d'un site

Si score allèle 1 < score allèle 2 => apparition d'un site

Schémas

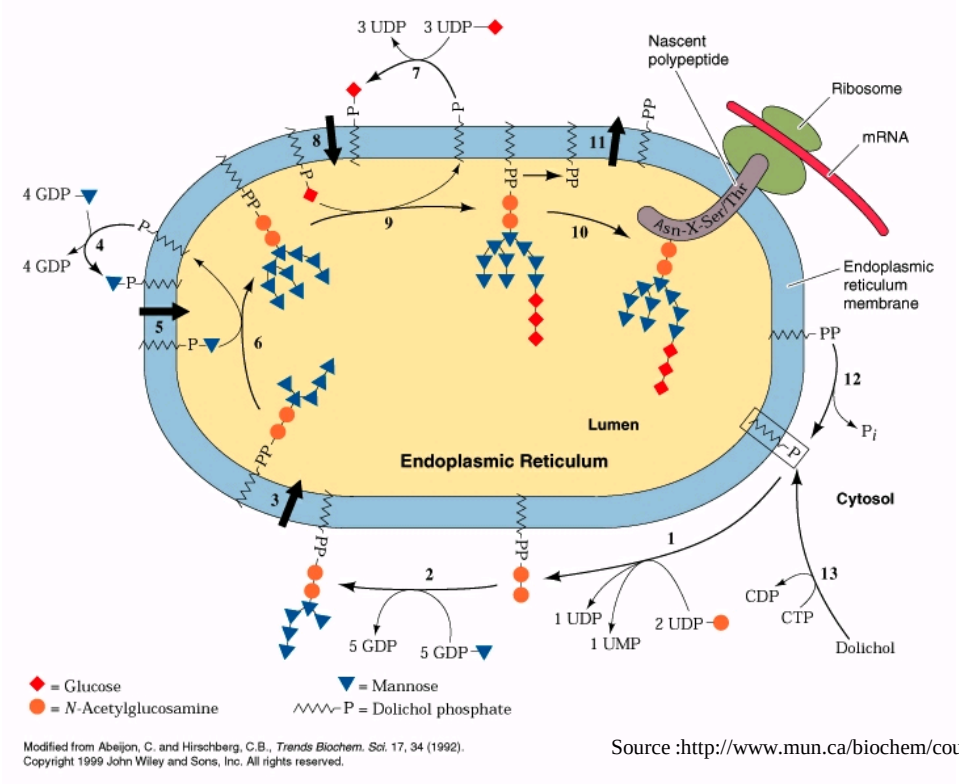
Phosphorylation :



Source: http://kinasephos.mbc.nctu.edu.tw/image/phosphorylation.jpg

Schémas

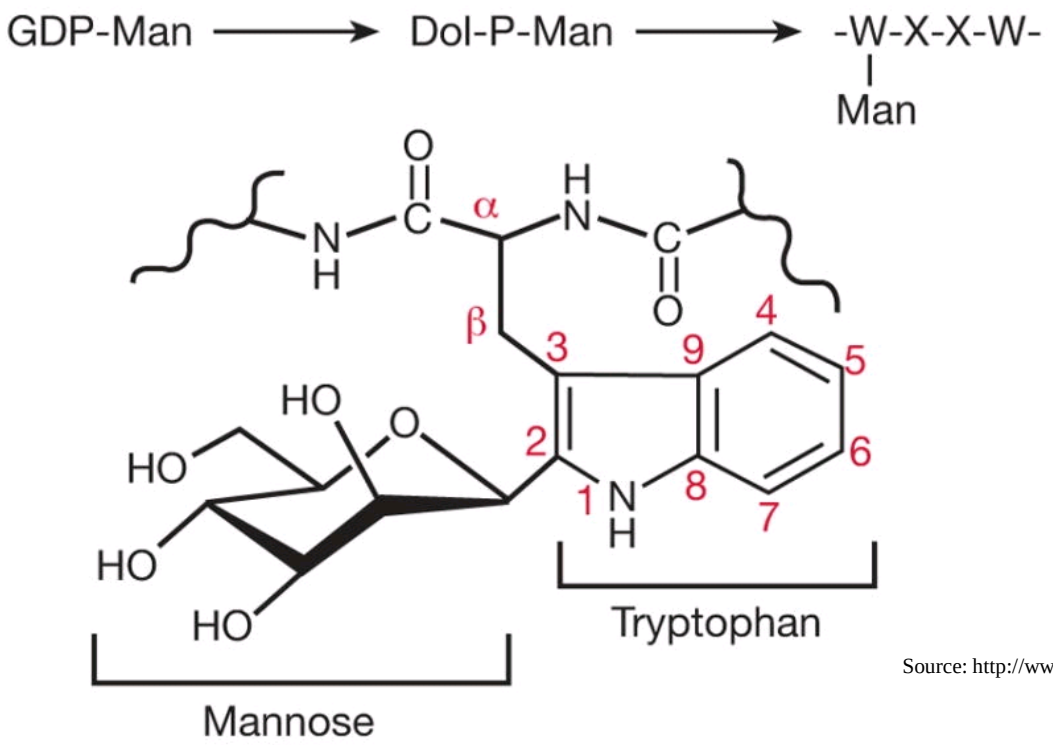
Glycosylation :



Source :http://www.mun.ca/biochem/courses/3107/images/VVP/Ch15/15-34.jpg

Schémas

Mannosylation :



Source: http://www.ncbi.nlm.nih.gov/books/NBK1947/bin/ch12f7.jpg

Schémas

Glypiation :

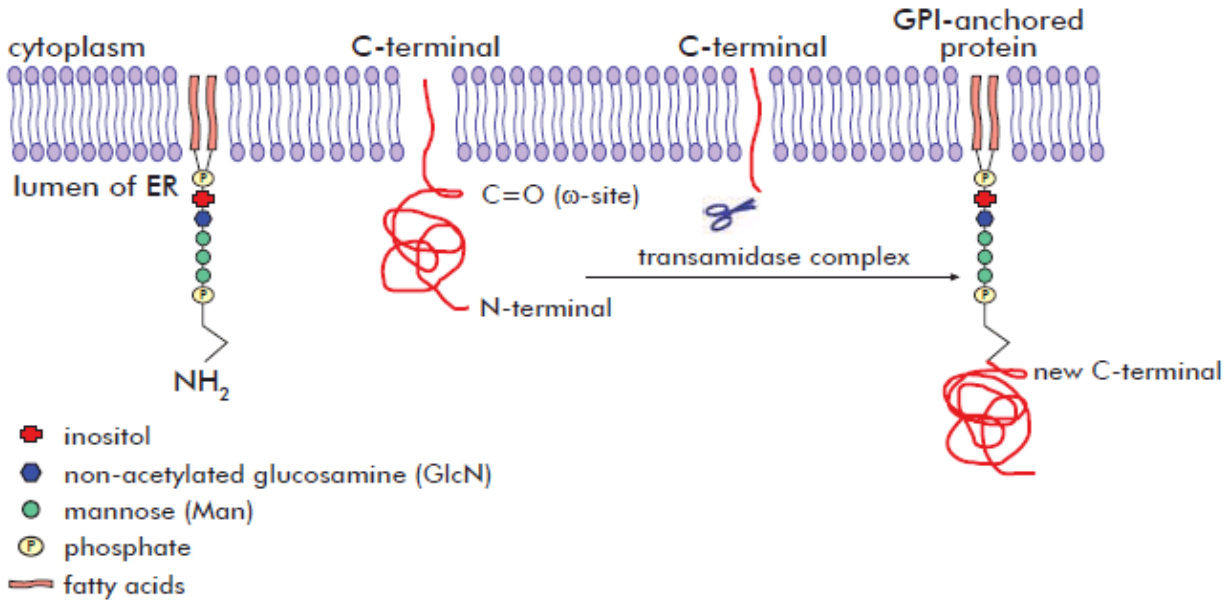


Figure 2. Schematic representation of the steps involved in the GPI-anchor event.
GPI: glycosylphosphatidylinositol.

Source : http://scielo.sld.cu/img/revistas/bta/v28n1/f0202111.gif

Description du pipeline

Chaque module fait appel à un programme de prédiction.

Les prédictions de perte / gain se font en comparant les résultats du programme par allèle:
- La perte « loss » d'une modification post-traductionnelle: allèle 1 avec signal / allèle 2 sans signal.
- Le gain « gain » d'une modification post-traductionnelle: allèle 2 avec signal / allèle 1 sans signal.

A cette comparaison s'ajoute une comparaison des scores (utilisé si absence de signal):
L'utilisateur fixe un seuil minimal pour le score de prédiction du programme (ex: 0.5).
Si la différence de score entre les 2 allèles >= au seuil minimum fixé en paramètre:
« perte? » ou « gain? ».

Spécification du pipeline:
- SNPs avec plus d'1 allèle alternatif (les résultats de chaque allèle alternatif sont comparés avec la référence et le fichier final contiendra 1 ligne par allèle alternatif)
- Indels traités pour les modules concernants les SNP codants non synonymes uniquement
- Formats d'entrée de variant_effect_predictor utilisés (.vcf, rsID, ensembl defaut)

Description du pipeline

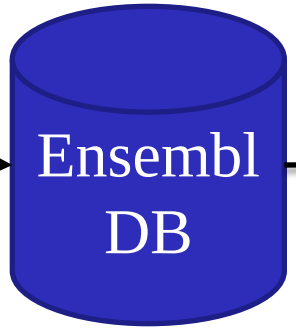
Données d'entrée

Liste de rsID

rs43580136
rs43697393
rs43625167
…

Format requis par « variant_predictor_effect.pl »

```
Test1   chr13   66779107    66779107    C/T 1
Test2   chr8    110793207   110793207   G/A 1
Test3   chr14   65037563    65037563    C/G 1
Test4   chr7    14850002    14850002    G/T 1
Test5   chr21   15882213    15882213    G/A 1
Test6   chr9    63041033    63041033    G/A 1
```



Source: personnelle

Registry file

```
use Bio::EnsEMBL::DBSQL::DBAdaptor;
use Bio::EnsEMBL::Variation::DBSQL::DBAdaptor;
use Bio::EnsEMBL::Registry;

Bio::EnsEMBL::DBSQL::DBAdaptor->new(
                '-species' => "Homo_sapiens",
                '-group' => "core",
                '-port' => 5306, '-host' => 'ensembldb.ensembl.org',
                '-user' => 'anonymous', '-pass' => "",
                '-dbname' => 'homo_sapiens_core_69_37' );

Bio::EnsEMBL::Variation::DBSQL::DBAdaptor->new(
                '-species' => "Homo_sapiens",
                '-group' => "variation",
                '-port' => 5306,
                '-host' => 'ensembldb.ensembl.org',
                '-user' => 'anonymous', '-pass' => '',
                '-dbname' => 'homo_sapiens_variation_69_37' );

Bio::EnsEMBL::Registry->add_alias("Homo_sapiens","human");
```

ensembl.registry contient toute les informations de connexion à votre base de données locale.

On y précise également l'espèce sur laquelle on travaille

http://www.ensembl.org/info/docs/variation/vep/vep_script.html#hgvs