Manuel de l'utilisateur de Wapam

Table des matières

Index des illustrations

1 Wapam, une recherche de motifs par automates pondérés

Wapam est un outil de recherche de motifs développé au sein de l'équipe de recherche SYMBIOSE et mis en ligne sur le site de la plate-forme OUEST-genopole®. Wapam peut rechercher rapidement des motifs protéiques ou nucléiques, avec ou sans erreur(s), dans des génomes complets, dans des banques de données et dans des banques personnelles (maxi. 80M).

L'interface Web permet aux utilisateurs de lancer leur requête sur le cluster de machines (genocluster) mis à disposition par la plate-forme ou d'utiliser l'accélérateur Rdisk. Rdisk est une architecture spécialisée conçu par l'équipe de recherche SYMBIOSE pour réduire considérablement le temps de recherche du motif dans les séquences cibles.

La première particularité de Wapam, est qu'il recherche des motifs exprimés en automates pondérés (WFA) (voir le chapitre 4). Les automates pondérés peuvent être générés à partir de motifs Prosite . Chaque séquence est enfilée progressivement dans cet automate. Il en ressort un score seuil qui permet d'évaluer l'adéquation de la séquence avec le motif. . Typiquement, un score simple : c'est le nombre d'erreurs de substitutions par rapport à un motif Prosite. Si le score passe au-dessus d'un certain seuil, le motif est détecté à la position courante (exemple : si une seule substitution est tolérée le score seuil sera égale à -1 et le motif sera détecté si le score est supérieur ou égal à -1. **Une recherche avec Wapam avec ou sans erreurs prend le même temps d'exécution.**

L'autre particularité de Wapam est son couplage avec la machine prototype Rdisk qui permet une **accélération matérielle** du calcul. Lors d'une étape de compilation, l'automate du motif est transformé en circuit spécialisé. Chacun des 31 processeurs reconfigurables qui composent Rdisk sont ensuite paramétrés avec ce circuits. La séquence est divisée en 31 morceaux qui sont traités dans chacun des processeurs.

2 Tutoriel : un exemple simple d'utilisation

Utilisation avec Rdisk

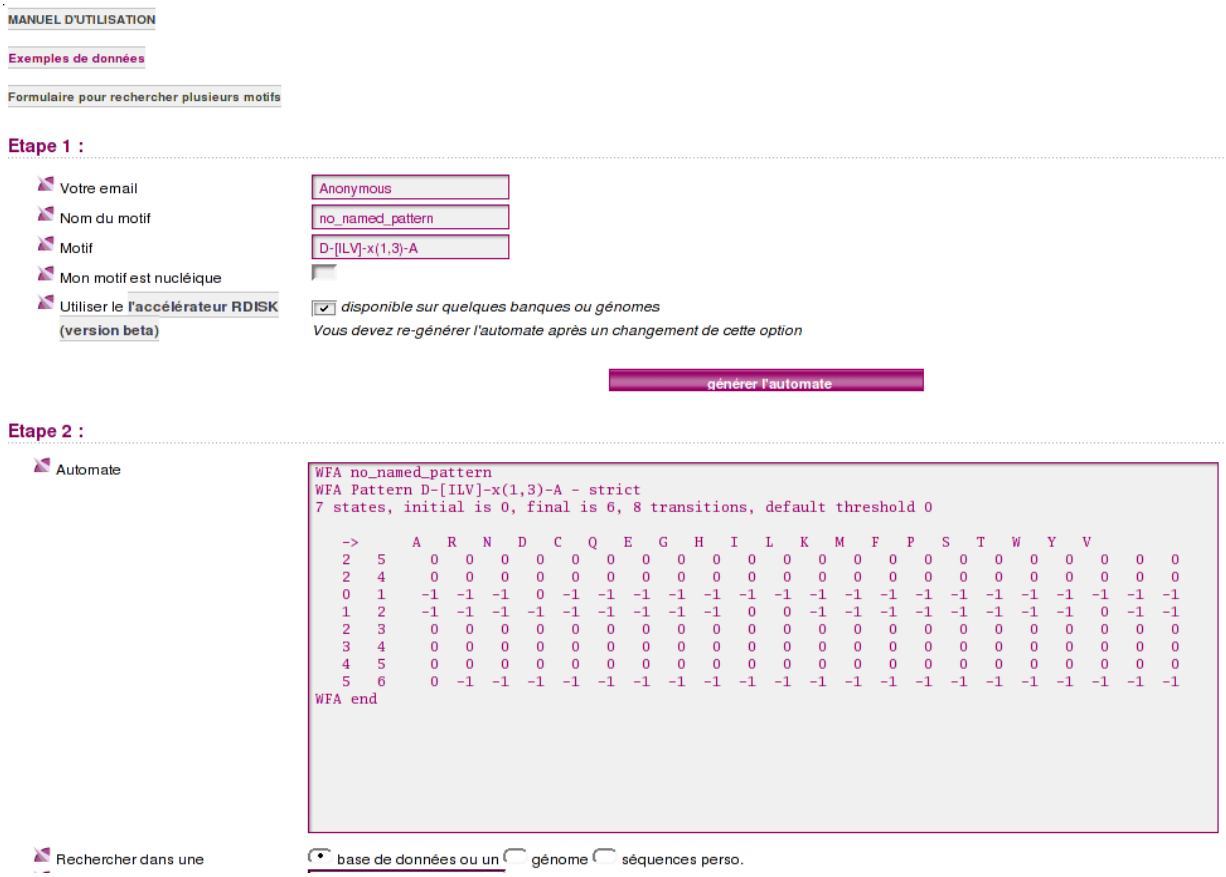


Illustration 1: exemple de saisie de données dans l'interface web

Nous souhaitons rechercher le motif Prosite D-[ILV]-x(1,3)-A dans la bases de données protéiques SwissProt. Il faut alors **générer l'automate** en appuyant sur le bouton correspondant. L'automate représentant ce motif se trouve dans l'illustration 1 Il est important de noter que si une modification de paramètres est effectuée alors que l'automate est généré, il faut le générer une nouvelle fois. Ici nous avons choisi d'utiliser Rdisk.

Illustration 1: Exemple de saisie de données

Il est possible de modifier l'automate, par exemple pour donner plus de poids à une transition (voir le chapitre 4).

Nous aurions pu aussi choisir de rechercher ce motif dans un génome comme dans l'illustration 2 Dans ce cas il faut préciser l'organisme et le ou les chromosome(s) – vous pouvez sélectionner plusieurs chromosomes avec la touche *maj.* - et s'assurer que l'option « each sequence matched » est sélectionnée.

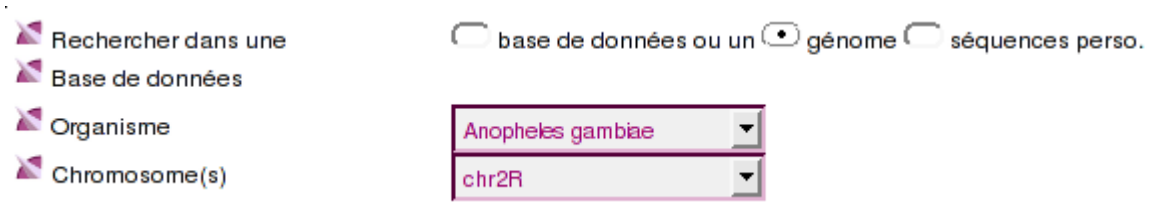


Illustration 2: Autre exemple de saisie de données dans l'interface web : recherche dans un génome

Enfin il est possible de réaliser une recherche de motif dans une banque personnelle (Illustration 3 Attention votre banque ne doit pas dépasser 80M et les séquences sont au format FASTA.

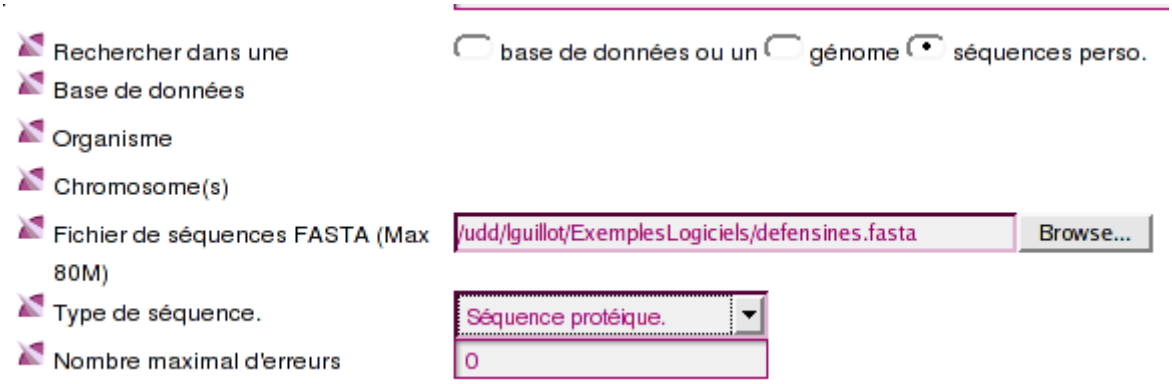
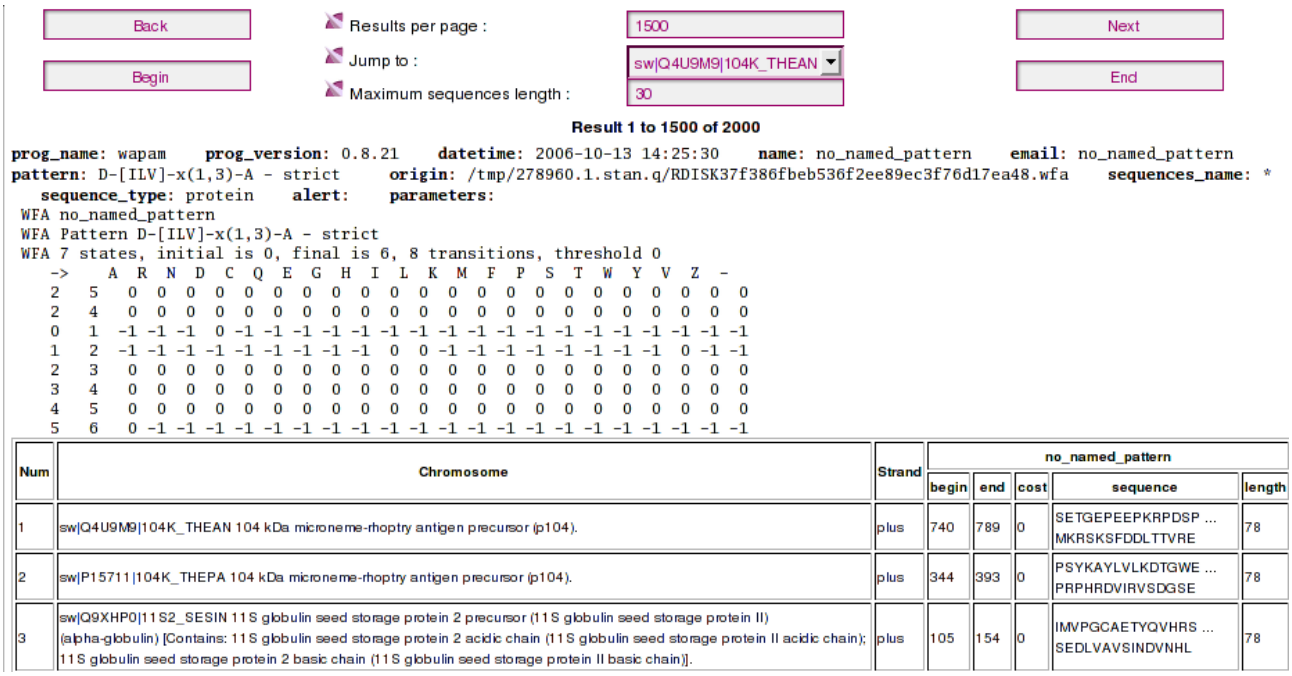


Illustration 3: Autre exemple de saisie de données dans l'interface web : recherche dans une banque personnelle.

Une page de mise en attente affiche un indicateur de progression de compilation et de passage des séquences comme indiqué sur l'illustration 4.

Illustration 4: Progression de la compilation des processeurs de Rdisk (FPGA) avant la filtration des séquences.

Les résultats sont alors ceux représentés dans l'illustration 5.



Back | Begin | Results per page : 1500 | Jump to : sw|Q4U9M9|104K_THEAN | Maximum sequences length : 30 | Next | End

Result 1 to 1500 of 2000

```
prog_name: wapam    prog_version: 0.8.21    datetime: 2006-10-13 14:25:30    name: no_named_pattern    email: no_named_pattern
pattern: D-[ILV]-x(1,3)-A - strict    origin: /tmp/278960.1.stan.q/RDISK37f386fbeb536f2ee89ec3f76d17ea48.wfa    sequences_name: *
   sequence_type: protein    alert:    parameters:
 WFA no_named_pattern
 WFA Pattern D-[ILV]-x(1,3)-A - strict
 WFA 7 states, initial is 0, final is 6, 8 transitions, threshold 0
    ->    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  Z  -
    2   5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    2   4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    0   1  -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
    1   2  -1 -1 -1 -1 -1 -1 -1 -1 -1  0  0 -1 -1 -1 -1 -1 -1 -1  0 -1 -1
    2   3   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    3   4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    4   5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    5   6   0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Num	Chromosome	Strand	no_named_pattern				
			begin	end	cost	sequence	length
1	sw|Q4U9M9|104K_THEAN 104 kDa microneme-rhoptry antigen precursor (p104).	plus	740	789	0	SETGEPEEPKRPDSP ... MKRSKSFDDLTTVRE	78
2	sw|P15711|104K_THEPA 104 kDa microneme-rhoptry antigen precursor (p104).	plus	344	393	0	PSYKAYLVLKDTGWE ... PRPHRDVIRVSDGSE	78
3	sw|Q9XHP0|11S2_SESIN 11S globulin seed storage protein 2 precursor (11S globulin seed storage protein II) (alpha-globulin) [Contains: 11S globulin seed storage protein 2 acidic chain (11S globulin seed storage protein II acidic chain); 11S globulin seed storage protein 2 basic chain (11S globulin seed storage protein II basic chain)].	plus	105	154	0	IMVPGCAETYQVHRS ... SEDLVAVSINDVNHL	78

Illustration *5* : Affichage des résultats de l'exemple en HTML..

Utilisation sans Rdisk

Les saisies sont les mêmes que dans l'illustration 1, il suffit de ne pas cocher la case Rdisk. L'illustration 6, montre le nombre de jobs en attente sur genocluster. La requête est placée dans cette file d'attente avant d'être exécutée sur un des noeuds du cluster de machines.

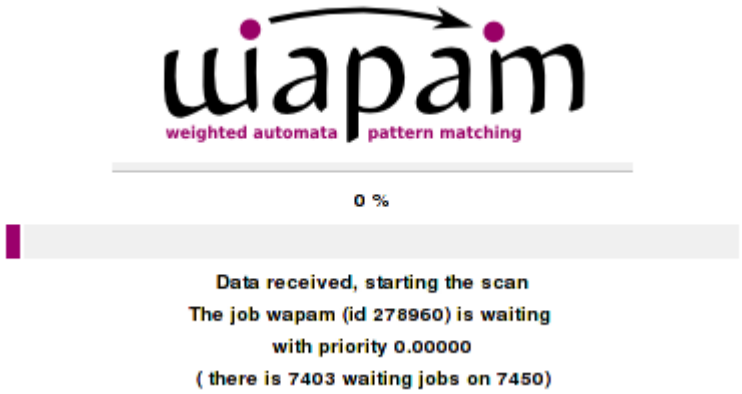


Illustration 6: Positionnement du job lancé dans la file d'attente des tâches de genocluster.

Les résultats sont les mêmes que dans l'illustration 5.

3 Entrée et sorties de données de Wapam

Données en entrée

Les paramètres à remplir sur le formulaire Web sont les suivants :

- Donner son **email** est optionnel, mais conseillé. Certaines recherches peuvent être assez longues, vous risquez donc de fermer votre navigateur et ainsi de perdre le lien sur la page résultat. Dans tous les cas le fichier résultat est sauvegardé 5 jours sur nos serveurs.
- Le **nom de motif** est également facultatif. Il vous permet de différencier vos requêtes lorsque vous en lancez plusieurs.
- Si votre motif est nucléique vous devez le préciser.
- Choisir d'utiliser **Rdisk** ou non. La machine spécialisée Rdisk permet d'accélérer les calculs (voir ci-dessous). C'est un prototype de recherche qui peut être souvent hors-service.
- Définir les **séquences cibles**. La plate-forme met à disposition environ 200 génomes et une vingtaine de banques de données. Des génomes et bases de données peuvent être rajoutées à la demande (webmaster@genouest.org). Si on utilise Rdisk ce choix est beaucoup plus limité, mais là encore nous pouvons faire des rajouts à la demande.Vous pouvez également importer vos séquences personnelles.
- Choisir le type de résultat : toutes les occurences de motifs (« **chaque match »**) ou juste les séquences qui matchent avec le motif **(« chaque sequence qui match »)**. Habituellement, vous choisirez « eatch match » (en particulier lorsque la recherche se fait dans un génome).

Formulaire pour rechercher plusieurs motifs

Accessible par un lien qui est en haut à gauche du formulaire. Il permet de lancer Wapam itérativement sur un ensemble de motifs (ensemble de motifs dans un format texte et non au format Word). Les autres paramètres d'entrée sont identiques.

Dans ce cas d'utilisation :

- les matrices des motifs ne sont pas modifiables manuellement.
- Les résultats sont exclusivement envoyés par mail : soit un mail par résultat, soit un seul mail pour tous les résultats. Les résultats sont alors enregistrés dans un unique fichier.

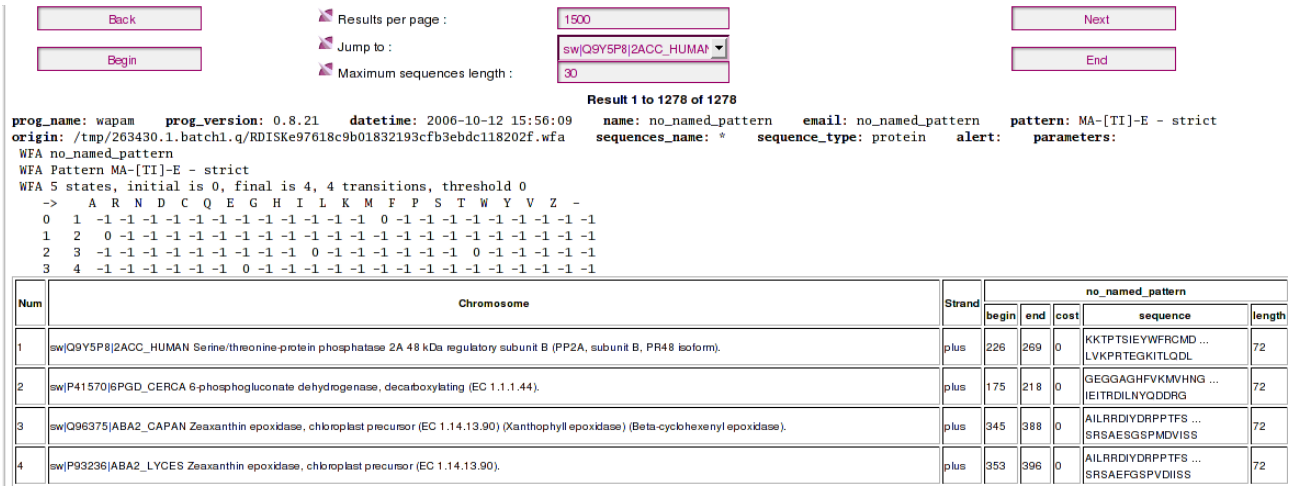
Pour avoir des renseignements sur le lancement d'un ensemble de motifs ou pour mettre en place un traitement avec de nombreux motifs, contactez webmaster@genouest.org.

Données de sortie

Les 3 formats de description des résultats contiennent exactement les mêmes données mais elles sont présentées différemment.

Format Web (HTML)

Le format HTML vous permet de visualiser vos données dans un tableau dans votre navigateur internet (illustration 7).



Back | Begin | Results per page : 1500 | Jump to : sw|Q9Y5P8|2ACC_HUMAN | Maximum sequences length : 30 | Next | End

Result 1 to 1278 of 1278

```
prog_name: wapam    prog_version: 0.8.21    datetime: 2006-10-12 15:56:09    name: no_named_pattern    email: no_named_pattern    pattern: MA-[TI]-E - strict
origin: /tmp/263430.1.batch1.q/RDISKe97618c9b01832193cfb3ebdc118202f.wfa    sequences_name: *    sequence_type: protein    alert:    parameters:
 WFA no_named_pattern
 WFA Pattern MA-[TI]-E - strict
 WFA 5 states, initial is 0, final is 4, 4 transitions, threshold 0
    ->    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  Z  -
    0   1  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1
    1   2   0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
    2   3  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1
    3   4  -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Num	Chromosome	Strand	no_named_pattern				
			begin	end	cost	sequence	length
1	sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform).	plus	226	269	0	KKTPTSIEYWFRCMD ... LVKPRTEGKITLQDL	72
2	sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44).	plus	175	218	0	GEGGAGHFVKMVHNG ... IEITRDILNYQDDRG	72
3	sw|Q96375|ABA2_CAPAN Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (Xanthophyll epoxidase) (Beta-cyclohexenyl epoxidase).	plus	345	388	0	AILRRDIYDRPPTFS ... SRSAESGSPMDVISS	72
4	sw|P93236|ABA2_LYCES Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).	plus	353	396	0	AILRRDIYDRPPTFS ... SRSAEFGSPVDIISS	72

Illustration 7: Exemple de la sortie HTML avec l'option « each sequence matched»

Le nombre de résultats affichés sur une page peut être déterminé en remplissant le champ texte « Result per pages » en haut de la page (par défaut 1500). Les données récupérées (illustration 7) sont :

- Le **nom du chromosome ou de la séquence**. Vous pouvez aller directectement au chromosome ou à la séquence qui vous intéresse en cliquant sur le champ « jump to » en haut de la page.
- Le **brin** (pour l'instant la recherche ne se fait que sur le brin plus)
- La **position de début et la position de fin** de la séquence affichée dans les résultats (et non celle du motif).
- le **coût** ou nombre d'erreurs par rapport au motif initial
- La **séquence** dont on peut sélectionner la longueur d'affichage dans le champ texte « maximum sequences length » en haut de la page (par défaut 30).
- La **longueur réelle** de la portion de la séquence affichée.

1996	sw|P13671|CO6_HUMAN Complement component C6 precursor.	plus	441	481	0	ISLIRGGRSEYGAAL ... RNIPCAVTKRNNLRK	69
1997	sw|P13671|CO6_HUMAN Complement component C6 precursor.	plus	553	593	0	EKQSPDYKSNAVDGQ ... QEEDCTFSIMENNGQ	69

Illustration 8: Exemple de résultats avec l'option « each match ». Dans cette séquence, le motif apparaît deux fois aux positions 481 et 593.

Format XML

Le format XML (illustration 9) est un format standard (cf http://www.w3.org/XML/1999/XML-in-10-points.fr.html) permettant d'enregistrer des données de façon à ce qu'elle puissent être relues facilement par un humain ou un programme. Vous en aurez peut être besoin si vous souhaitez traiter les données automatiquement par un script que vous souhaitez écrire vous même. En réalité, le format Web est produit à partir du format XML.

```
<?xml version="1.0" encoding="UTF-8" ?>
<result  prog_name="wapam"  prog_version="0.8.21"  datetime="2006-10-12 15:56:09"  name="no_named_pattern"  email="no_named_pattern"  pattern="MA-[TI]-E - stri
  <occurrence sequence="sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform)." complement="
    <pattern name="no_named_pattern" begin="226" end="269" cost="0"  >
      <sequence type="protein"  >
        KKTPTSIEYWFRCMDLDGDGALSMFELEYFYEEQCRRLDSMAIEALPFQDCLCQMLDLVKPRTEGKITLQDL
      </sequence>
    </pattern>
  </occurrence>
  <occurrence sequence="sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)." complement="plus"  >
    <pattern name="no_named_pattern" begin="175" end="218" cost="0"  >
      <sequence type="protein"  >
```

Illustration 9: Exemple de la sortie au format XML

Format CSV

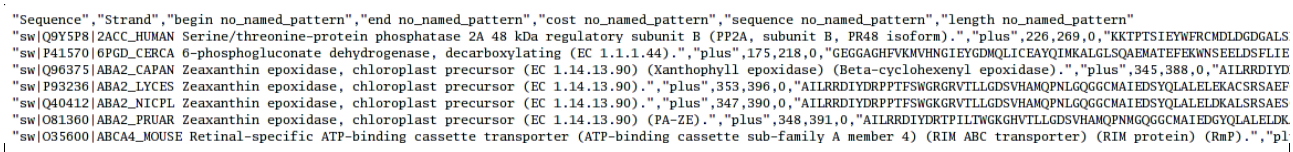
Le format CVS (illustration 10) permet d'importer vos données dans n'importe quel logiciel tableur comme Excel ou Open Office.Calc. Il est lui aussi traduit à partir du format XML. Le format CSV utilisé par WAPAM est le suivant :

- le séparateur de champs est la virgule,
- le séparateur de texte est le guillemet.

Pour récupérer un document CSV dans Excel,

1. Sur l'interface web de WAPAM, cliquez sur le bouton droit de la souris sur le lien 'Description des résultats au format CSV'', enfin cliquez sur 'Enregistrer la cible du lien sous ...'
2. Dans Excel : Fichier/Ouvrir
3. Sélectionnez "tous" dans 'type de fichier'
4. Sélectionnez le type de fichier CSV et validez
5. Sélectionnez toute la colonne A

6. Dans le menu "Données" sélectionnez "Convertir"
7. Choisissez l'option "délimité" et appuyez sur "suivant"
8. Indiquez comme séparateur la virgule et comme indicateur de texte le guillemet
9. Cliquez sur terminer
10. Vous n'avez plus qu'à formater votre tableau comme bon vous semble.



```
"Sequence","Strand","begin no_named_pattern","end no_named_pattern","cost no_named_pattern","sequence no_named_pattern","length no_named_pattern"
"sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform).","plus",226,269,0,"KKTPTSIEYWFRCMDLDGDGALS
"sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44).","plus",175,218,0,"GEGGAGHFVKMVHNGIEYGDMQLICEAYQIMKALGLSQAEMATEFEKWNSEELDSFLIE
"sw|Q96375|ABA2_CAPAN Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (Xanthophyll epoxidase) (Beta-cyclohexenyl epoxidase).","plus",345,388,0,"AILRRDIYD
"sw|P93236|ABA2_LYCES Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).","plus",353,396,0,"AILRRDIYDRPPTFSWGRGRVTLLGDSVHAMQPNLGQGGCMAIEDSYQLALELEKACSRSAEF
"sw|Q40412|ABA2_NICPL Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).","plus",347,390,0,"AILRRDIYDRPPTFSWGKGRVTLLGDSVHAMQPNLGQGGCMAIEDSYQLALELDKALSRSAES
"sw|O81360|ABA2_PRUAR Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (PA-ZE).","plus",348,391,0,"AILRRDIYDRTPILTWGKGHVTLLGDSVHAMQPNMGQGGCMAIEDGYQLALELDK
"sw|O35600|ABCA4_MOUSE Retinal-specific ATP-binding cassette transporter (ATP-binding cassette sub-family A member 4) (RIM ABC transporter) (RIM protein) (RmP).","pl
```

Illustration 10 Exemple de la sortie au format CVS

Remarque sur le nombre de résultats

Nous avons limité le nombre de résultats en sortie de Wapam (par genocluster : 2000 / par Rdisk : 500). En effet, une requête avec un trop grand nombre de réponses apparaît difficilement interprétable : il est alors préférable que l'utilisateur biologiste d'affine sa recherche. Il est toutefois possible d'augmenter ces seuils en contactant webmaster@genouest.org.

4 Quelques détails sur le fonctionnement de Wapam

Les automates pondérés (WFA)

Un automate caractérisant un motif sera représenté par l'ensemble des positions du motif, reliés entre elles par des transitions (illustration 11). L'automate est pondéré, c'est à dire que chaque transition est étiquetée par une lettre qui peut être lue selon l'alphabet de la séquence (bases nucléique ou protéique) et par un poids.

La séquence est progressivement « enfilée » dans l'automate, et, à chaque position, le poids de sa transition est additionné au score. Ce poids reflète l'adéquation d'une partie de la séquence cible (banque ou génome) avec la lettre lue à cette position dans le motif. Par défaut ce poids est égal à -1 si la lettre n'est pas la même (substitution) et à 0 si c'est la même.

Le motif est reconnu lorsque l'état final est actif avec un score supérieur ou égal au score ou seuil d'erreur fixé. Par exemple si une erreur est tolérée le seuil sera égal à -1.

Sur l'illustration 11 présentant un exemple d'automate pondéré, chaque rond est un état, chaque flèche est une transition.

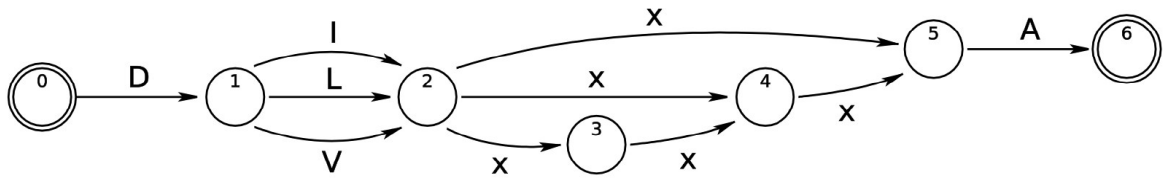


Illustration 11 : *un automate pondéré du motif D-[ILV]-x(1,3)-A.*

Les automates utilisés par Wapam sont sous la forme suivante (illustration 12). Par exemple, si la portion de séquence qui passe dans l'automate passe de l'état 0 à 1 en lisant un D le coût sera de 0 sinon le coût sera de -1.

```
7 states, initial is 0, final is 6, 8 transitions, default threshold 0

  ->      A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
  2   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  0   1  -1  -1  -1   0  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
  1   2  -1  -1  -1  -1  -1  -1  -1  -1  -1   0   0  -1  -1  -1  -1  -1  -1  -1   0  -1  -1
  2   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  3   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  4   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  5   6   0  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
```

Illustration 12 : *Exemple d'automate représentant un motif Prosite : D-[ILV]-x(1,3)-A.*

Les poids peuvent être plus généraux que le simple décompte « 0 / -1 » ; il est possible de modifier manuellement l'automate. Par exemple la substitution de D par N, R ou A en première position peut coûter -3 au lieu de -1 (Illustration 13).

```
->      A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
2   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
2   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   1  -3  -3  -3   2  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
1   2  -1  -1  -1  -1  -1  -1  -1  -1  -1   0   0  -1  -1  -1  -1  -1  -1  -1   0  -1
2   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
3   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
4   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
5   6   0  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
```

Illustration 13 : *Exemple d'automate modifié à la main.*

La plateforme dispose d'autres outils pour générer des automates pondérés (génération de poids « à la BLOSSUM », utilisation de matrices poids/position PWM...) Contactez webmaster@genouest.org pour des questions à ce sujet.

Wapam et Wapam/Rdisk

Wapam peut être utilisé de deux façons (Illustration 14) : soit il est lancé sur genocluster (comme tous les autres logiciels de la plate-forme) et la recherche se fait sur un noeud du cluster, soit il est couplé avec l'architecture Rdisk qui parallélise la recherche sur un ensemble de cartes.

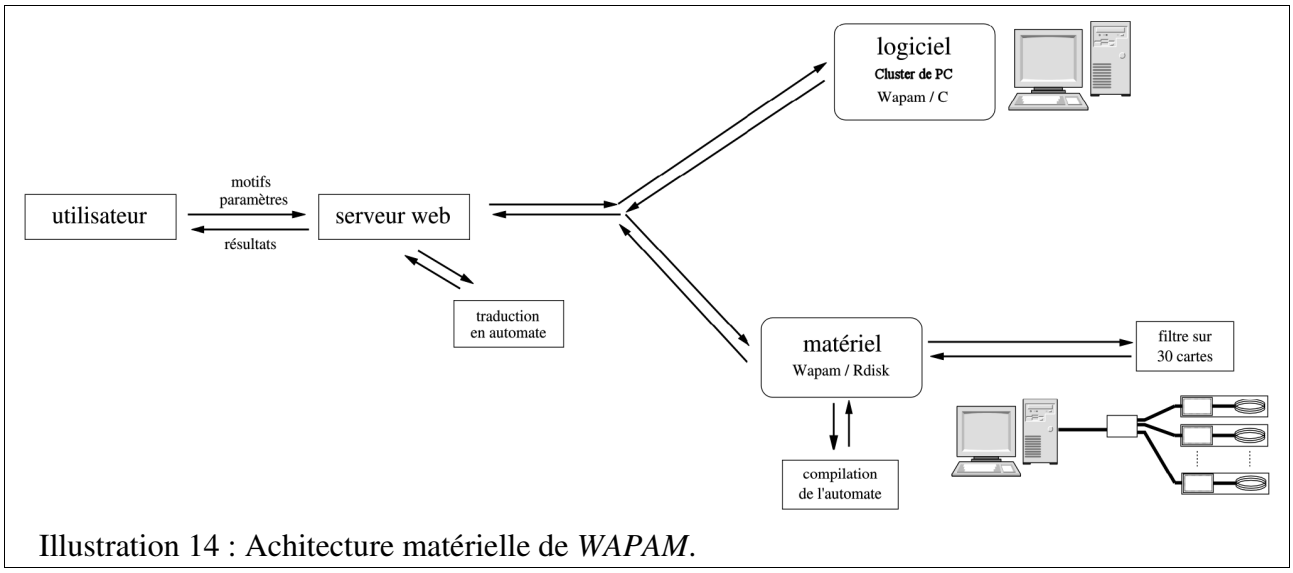


Illustration 14 : Achitecture matérielle de *WAPAM*.

Rdisk est une architecture spécialisée constituée de plusieurs dizaines de cartes (actuellement 31). Chaque carte contient un processeur reconfigurable (FPGA) couplé à un disque dur. Les automates pondérés sont directement cablés sur les FPGA, ce qui permet une évaluation simultanée des états.

Ce cablage utilise autant d'élements matériels que de transitions d'états dans l'automate. Les processeurs utilisés ont une surface pouvant cabler des automates ayant jusqu'à une centaine de transitions. Les 31 cartes se partagent le balayage de la banque ou du génome (1/31^{ieme} par carte). L'ensemble du prototype Rdisk a été conçu pour filtrer rapidement les bases de données, les disques durs étant directement reliés aux processeurs FPGA.

Rdisk étant un prototype de recherche, il n'est pas toujours en service. Si vous avez besoins de calculs intensif en recherche de motifs, contactez la plateforme (webmaster@genouest.org) pour que nous mettions en place un traitement adapté de vos données ou de vos motifs.

Performances

L'illustration15 présente une comparaison des temps de recherche de motif entre l'implémentation logicielle de Wapam et l'accélération matérielle Wapam/Rdisk (moyenne sur 50 motifs pris aléatoirement parmi un ensemble de 3331 motifs). Pour ne pas surcharger les serveurs, la recherche peut être arrêtée dès qu'il y a plus d'un certain nombre de résultats (auto-stop).

Dans tous les cas, une recherche avec Wapam avec ou sans erreurs prend le même temps d'exécution. Sur la version logicielle, le temps d'exécution est linéaire par rapport à la taille de l'automate (et donc du motif). Pour Wapam/Rdisk, tous les motifs sont traités dans le même temps (tant qu'ils sont acceptés par Rdisk, c'est-à-dire tant qu'il n'y a pas plus qu'une centaine de transitions).

	Wapam logiciel	***Wapam + autostop 2000***	***Wapam/Rdisk***	***Wapam/Rdisk + précompilation***
1 motif	2605 s	2003 s	72 s	23 s
3331 motifs	100 jours*	77 jours*	< 3jours	< 1jour

Illustration 15 : *Comparaison des temps de recherche de motif (* : estimations)*

L'accélération apportée par Rdisk est encore plus importante à partir du deuxième lancement, lorsque les motifs ont déjà été compilés, car Wapam / Rdisk se souvient des automates pondérés compilés précedemment. La modification du seuil d'erreur ne demande pas une nouvelle compilation.

Besoins spécifiques

Nous sommes à votre disposition (webmaster@genouest.org) pour collaborer sur des tâches particulières, comme par exemple :

- ajouter d'autres banques de données,
- réaliser des automates pondérés répondant à des objectifs particuliers,
- mettre en place sur le cluster ou sur Rdisk des calculs intensifs (grand nombre de séquences, de motifs/d'automates, lancements itérés, analyse de résultats...); nous pouvons paramétrer finement Wapam pour obtenir les meilleurs temps de calculs sur votre application,
- vous fournir un accès à Wapam en ligne de commande sur genocluster

Références

Merci de citer la référence suivante dans vos travaux utilisant Wapam.

Stéphane Guyetant, Mathieu Giraud, Ludovic L'Hours, Steven Derrien, Stéphane Rubini, Dominique Lavenier, and Frédéric Raimbault. Cluster of re-configurable nodes for scanning large genomic banks. *Parallel Computing*, 31(1):73-96, 2005.