concevoir des candidats médicaments sur internet (1/2)

outils et services web gratuits pour la chimie médicinale

publié le 15.12.20 |par bruno villoutreix

cet article en deux volets est consacré à la recherche de nouveaux principes actifs in silico (drug design in silico), en se limitant aux principes actifs constitués de petites molécules, à l'exclusion des vaccins et biomédicaments. le premier volet (ci-dessous), après quelques rappels sur les grandes étapes de la recherche d'un principe actif, présente les bases de données disponibles en libre accès regroupant cibles thérapeutiques et petites molécules chimiques susceptibles d'interagir avec ces cibles. le second volet traite du criblage virtuel en ligne et des deux principales approches de celui-ci : celle qui utilise les propriétés des petites molécules chimiques (en anglais, ligand-based virtual screening (lbvs)) et celle qui utilise la structure tridimensionnelle de la cible thérapeutique (en anglais, structure-based virtual screening (sbvs)).

1. introduction

il faut entre 12 et 15 ans et plus d'un milliard d'euros en moyenne pour développer un nouveau médicament. c'est un processus complexe qui implique de nombreuses étapes et des compétences variées. l'enjeu est de taille pour le chercheur de médicament : s'il a déjà aujourd'hui à sa disposition des données massives (« big data ») générées pour le moment par les technologies haut-débit, dans un futur proche il pourra compter sur d'autres données, comme celles provenant des objets connectés.

les données actuelles proviennent pour la plupart des « -omiques » (la génomique, la protéomique…) et des criblages chimiques. traitées par certains algorithmes, elles devraient permettre de faire des progrès considérables dans le secteur de la santé. produites par des milliers de laboratoire de recherche dans le monde, elles sont stockées dans des « entrepôts numériques » qui se matérialisent sous forme de bases de données gratuites ou commerciales. il faut savoir qu'il existe actuellement plus de 1000 bases de données contenant des milliards d'informations dans le secteur santé au sens large, consultables *via* internet. si les données sont essentielles – considérées par beaucoup comme l'or noir du 21ème siècle – il faut évidemment parvenir à les manipuler et les analyser afin de leur donner du sens et développer des modèles prédictifs pertinents.

un des défis majeurs pour le cyber-chasseur de médicaments est de trouver les données et les logiciels de traitement nécessaires à l'aboutissement de son projet. il existe bien évidemment des logiciels commerciaux mais l'on trouve aussi dans le cyber-espace des milliers de logiciels gratuits. ces outils sont pour la plupart facilement accessibles depuis des serveurs en ligne *via* n'importe quel navigateur web moderne. depuis plus de

20 ans, je collecte dans la littérature scientifique des bases de données ouvertes dans le domaine du médicament et des logiciels gratuits permettant de manipuler ces données (villoutreix et col., drug discovery today, 2013, 18 :1081-9 ; singh, chaput et villoutreix, briefings in bioinformatics, 2020, sous presse). je présente le fruit de cette veille quotidienne sur un site internet que j'ai développé (www.vls3d.com), notamment sur la page « shortlist » qui répertorie les principaux outils. grâce aux informations extraites de ce site, nous allons explorer plusieurs services en ligne qui facilitent le design de candidats médicaments, depuis les bases de données dédiées aux cibles thérapeutiques jusqu'aux outils de criblage virtuel.

2. les grandes étapes du processus de recherche d'un nouveau médicament

développer un médicament est un processus long, coûteux et risqué. le taux d'échec est généralement très élevé en raison de l'immense complexité des systèmes biologiques et des mécanismes moléculaires impliqués dans les pathologies. dans les années 1980, les agents thérapeutiques étaient essentiellement classés en deux catégories : les vaccins et les médicaments de type petites molécules chimiques. ces dernières années, une nouvelle classe est apparue : les « biomédicaments ». ceux-ci incluent par exemple les anticorps monoclonaux, les protéines thérapeutiques recombinantes, certains peptides, la thérapie génique et la thérapie cellulaire (certains auteurs classent aussi les vaccins dans cette catégorie). ainsi actuellement, et pour simplifier, nous pouvons dire qu'il existe les biomédicaments, qui ont comme point commun le fait de faire appel à une source biologique comme matière première du principe actif et les petites molécules chimiques, dont le principe actif est généralement issu de la synthèse chimique, souvent inspirée de produits naturels. dans le cadre de cet article, nous nous limiterons aux approches concernant les médicaments de type petites molécules chimiques, mais il faut souligner qu'il existe de nombreuses autres approches dédiées aux biomédicaments.

comment trouver des petites molécules chimiques candidates médicaments ? dans le passé, la découverte de nouveaux médicaments résultait soit de la sérendipité, c'est-à-dire d'un hasard parfois heureux pour un chercheur, soit de l'utilisation de produits naturels. de nos jours, même si elle n'exclut pas les heureux hasards, la recherche de nouveaux médicaments est basée sur des méthodes rationnelles et structurées. ces méthodes sont encore loin d'être parfaites. de manière schématique (fig. 1), on peut découper le processus selon les grandes étapes suivantes (nb : ces étapes sont différentes dans le cas de criblage phénotypique, pour la recherche d'un biomédicament ou pour le repositionnement d'un médicament existant) :

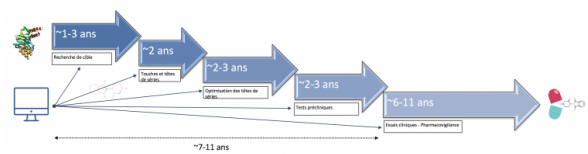

figure 1 - les grandes étapes du développement d'un nouveau médicament. différentes approches in silico peuvent intervenir le long du processus.

- **étape nº 1 : identification et sélection d'une cible ou de plusieurs cibles *a priori* impliquée(s) dans une pathologie. généralement, une cible thérapeutique est une protéine qui est découverte en utilisant plusieurs approches expérimentales (biologie moléculaire, génomique, méthodes biophysiques, criblage chimique…). cette étape est critique du processus de découverte d'un médicament.**
- **étape nº 2 : identification de petites molécules (dites « touches » ou « hits ») qui interagissent avec la cible ou les cibles (par exemple une petite molécule qui bloque le site catalytique d'une protéine surexprimée dans un cancer). cette étape implique des approches de criblage expérimental et/ou virtuel.**
- **étape nº 3 : optimisation des touches vers les têtes de séries (molécules plus efficaces que la touche). cette étape va impliquer plusieurs cycles de chimie médicinale, l'utilisation d'approches biophysiques, de modélisation informatique…**
- **étape nº 4 : optimisation des têtes de séries vers le candidat médicament (molécules encore plus efficaces et plus sûres). de multiples cycles de chimie médicinale, de tests expérimentaux, biophysiques… et des prédictions virtuelles sur ordinateur dites « in silico » seront nécessaires pour identifier ces molécules.**
- **étape nº 5 : tests précliniques. ces études permettent d'acquérir les premières connaissances indispensables sur le comportement d'un candidat médicament avant les essais chez l'homme. les expérimentations sont essentiellement menées sur l'animal mais plusieurs algorithmes peuvent aussi guider les travaux.**
- **étape nº6 : les essais cliniques sur l'homme sont utilisés pour démontrer l'efficacité des molécules développées avant la mise sur le marché. il y a plusieurs phases :**
  - **dans la phase i, le nouveau traitement ou vaccin est généralement administré à un petit groupe de volontaires en bonne santé.**
  - **si la phase i a donné des résultats probants, une autorisation est demandée pour réaliser un essai auprès d'un plus grand groupe de volontaires. les essais de phase ii incluent généralement des patients malades. à ce stade, la performance du médicament peut aussi être comparée à celle d'un placebo administré à un autre groupe de patients.**
  - **si les résultats de la phase ii sont encourageants, la phase iii est initiée. cet essai est mené à plus grande échelle et inclut souvent plusieurs centaines de volontaires originaires de différents pays. il s'agira entre autres de démontrer l'innocuité et l'efficacité du nouveau médicament ou vaccin. ces étapes sont généralement suivies d'une phase de pharmacovigilance, après la commercialisation des produits, afin de repérer d'éventuels effets indésirables non détectés durant les étapes précédentes.**

de nos jours, plusieurs technologies et disciplines innovantes, comme la bioinformatique [1] et la chémoinformatique[2], arrivent en renfort des approches expérimentales pour sans cesse améliorer l'efficience des premières étapes du processus (hillisch et col., chemmedchem, 2015, 10 : 1958-1962). ces outils informatiques peuvent utiliser le « big data » et certaines approches d'intelligence artificielle encore actuellement de « bas niveau » (c'est-à-dire généralement des méthodes d'apprentissages automatiques, supervisées ou non, qui apprennent à partir des données injectées dans le système), des approches de simulation moléculaire... certains logiciels peuvent aussi aider au moment des essais cliniques.

3. des cibles thérapeutiques et des molécules chimiques en « open access » dans les bases de données

3.1. les bases dédiées aux cibles thérapeutiques

un nombre important de bases de données est dédié aux cibles (potentiellement) thérapeutiques. a titre d'exemple, supposons que nous recherchions des informations sur la coagulation sanguine. il est possible de visualiser cette cascade de réactions complexes sur le site de la base reactome (table 1). par recherche avec le mot-clé « hemostasis » (en anglais), on obtient une première visualisation générale de ce système biologique (fig. 2). en cliquant sur l'image, on visite alors une nouvelle page web qui contient des informations plus précises. il est possible de sélectionner une étape, par exemple la formation du caillot sanguin (« formation of fibrin clot » en anglais) et de visualiser toutes les interactions connues, les protéines impliquées dans cette étape. l'interface est totalement interactive et on peut agrandir une image avec la souris de son ordinateur. dans l'interface de reactome, on peut cliquer sur une protéine de la cascade, par exemple le facteur xa[3], et l'on a directement un lien vers une autre base de données, la « protein data bank (pdb) » (ou vers la pdb europe), qui contient des informations sur la structure 3d expérimentale de cette protéine.

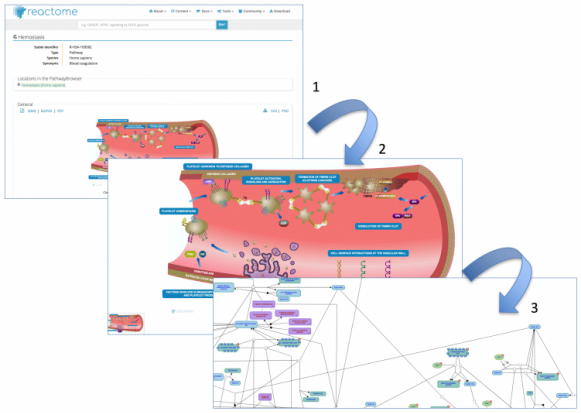

figure 2 - différents niveaux de visualisation de la cascade de coagulation sur la base reactome

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

dans les bases on trouve le nom facteur x ou facteur xa, le « a » indique que l'enzyme a été activée et est pleinement fonctionnelle. le numéro d'identification du facteur xa proposé dans l'interface reactome pour la pdb est « 4y6d », mais plusieurs autres structures 3d sont disponibles. si l'on va sur la pdb, avec comme code de recherche « 4y6d », on va retrouver le facteur xa ainsi que le fichier qui contient les coordonnées atomiques de sa structure. de nombreuses autres informations importantes sont disponibles sur le site. la protéine peut alors être visualisée de manière interactive directement sur le site de la pdb (fig. 3).

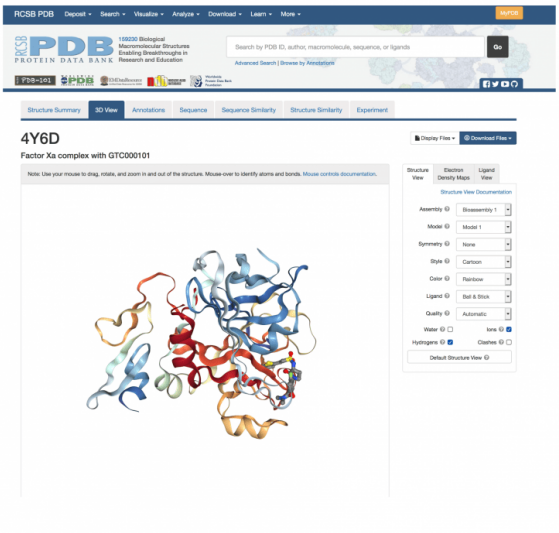

figure 3 - recherche sur la protein data bank de la structure 3d du facteur x, une protéine importante de la coagulation sanguine

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

l'existence d'une structure 3d d'une cible est une information critique pour le chasseur de médicaments. cela suggère qu'il sera possible de rechercher des petites molécules chimiques par des approches de criblage virtuel et de modélisation moléculaire[4]. ces informations vont aussi guider la synthèse chimique afin de construire des molécules ayant plus d'affinité pour la cible. il est évidemment plus simple de trouver une petite molécule qui bloque une poche catalytique d'une cible lorsqu'on peut voir la distribution des atomes dans l'espace, étudier les propriétés de la poche ou de la cavité qui va interagir avec cette petite molécule.

par ailleurs, l'interface reactome permet d'obtenir le code uniprot du facteur x (chez l'homme, p00742). il est alors possible d'aller chercher d'autres informations sur cette protéine dans la base uniprot (une base de données qui contient des millions de séquences de protéines et de nombreuses autres informations), par exemple rechercher s'il existe des patients ayant des mutations dans le gène qui code pour le facteur x. si la protéine qui nous intéresse n'a pas de structure 3d expérimentale (par cristallographie ou par rmn), il est souvent possible d'utiliser des approches de modélisation par homologie, qui visent à prédire la structure 3d d'une protéine en utilisant des patrons moléculaires (des protéines connues en 3d avec une identité de séquence proche de la séquence de la protéine que l'on cherche à prédire en 3d). plus de 40 millions de modèles structuraux sont accessibles sur swiss-model et modbase (table 1). il est à noter que des dizaines de services permettent de faire de la modélisation par homologie *via* le web. enfin, dans cette visite rapide des bases de données concernant les cibles, il faut signaler que certaines bases sont plus spécialisées sur l'aide à la sélection d'une cible. ces bases structurent l'information des cibles connues ou potentielles ainsi que leurs implications possibles dans des pathologies. par exemple, trois bases ont été publiées récemment sur cet axe : open targets, therapeutic target database (ttd) et pharos. sur open targets, en recherchant le facteur xa (*via* le mot-clé « fx »), l'on observe que 139 maladies sont associées à cette protéine (fig. 4). sur ttd il est possible de trouver les cibles qui sont actuellement concernées dans des essais cliniques en cours.

figure 4 - recherche d’informations sur une cible, le facteur x (fx), dans la base open targets.

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

3.2. les bases dédiées aux petites molécules chimiques

en ce qui concerne les petites molécules chimiques, ici aussi de très nombreuses bases de données sont accessibles sur internet. par exemple pubchem et chembl contiennent des millions de petites molécules chimiques annotées (table 1). ainsi sur pubchem, dans la section « bioassays », il est possible de rechercher des molécules qui touchent une cible, par exemple notre facteur xa. avec cette recherche dans le moteur de recherche de l’interface, on arrive par exemple sur une page qui montre des molécules testées sur cette protéine et les affinités expérimentales mesurées (fig. 5).

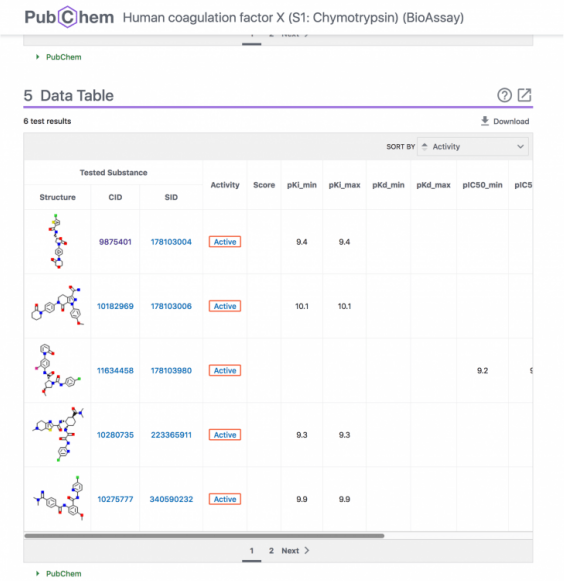

figure 5 - recherche de petites molécules qui inhibent le facteur x sur pubchem

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

il est possible d’aller ensuite visiter une base de données qui contient des médicaments déjà sur le marché ou en phases cliniques comme drugbank. on recherche des informations par mots-clés ou par structure chimique. ainsi en recherchant « coagulation factor xa » et « cible » (« target » en anglais), on trouve des liens vers uniprot ou la pdb, mais aussi une table qui montre les relations entre la protéine et les médicaments

existants (fig. 6).

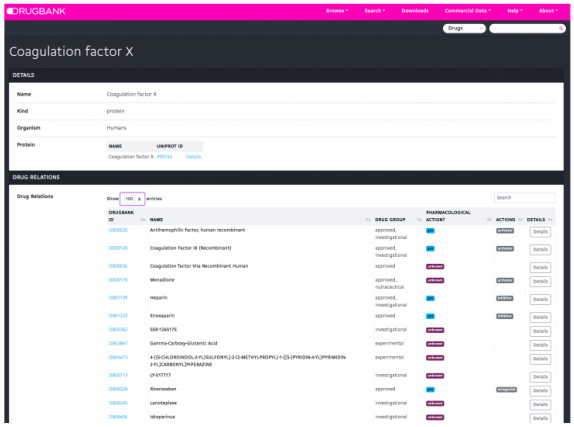

figure 6 - identification de médicaments qui agissent sur le facteur x dans drugbank.

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

on constate par exemple qu'il y a plusieurs médicaments déjà sur le marché qui agissent sur cette protéine, par exemple le rivaroxaban, un anticoagulant. un clic sur le nom de code de cette molécule dans la base (« identification number db06228 ») et de multiples informations apparaissent dans une nouvelle fenêtre. d'autres bases de données vont structurer des informations différentes : certaines peuvent être spécialisées en produits naturels (par exemple la base de données nubbe, extraits de plantes du brésil) ou encore en extraits de plantes utilisées en médecine traditionnelle chinoise (yatcm) ou indienne (imppat). il est généralement possible de télécharger les petites molécules présentes dans ces bases de données dans des formats électroniques spéciaux (par exemple formats sdf ou mol2 ou smiles). dès lors, il peut être intéressant d'analyser avec un logiciel approprié ces petites molécules sur son ordinateur. l'outil datawarrior est particulièrement approprié, il s'installe facilement sur tous les os et permet de faire de multiples calculs ainsi que de visualiser les données de diverses manières. il existe quelques tutoriels en français sur cet outil (par exemple sur radarweb création, voir table 1). avec datawarrior, il est aussi possible de télécharger automatiquement toutes les molécules présentes sur wikipédia[5]*via* le menu « database » si l'ordinateur est connecté à internet (fig. 7). plusieurs analyses statistiques sont alors réalisables pour étudier plus en profondeur certaines propriétés physico-chimiques des composés. par exemple, il est possible de calculer la masse moléculaire ou d'autres propriétés et de rechercher une valeur moyenne pour une propriété donnée. une analyse en composantes principales sur les propriétés choisies est réalisable, c'est une possibilité pour visualiser l'espace chimique. il est aussi possible de faire des recherches directement sur chembl avec cet outil.

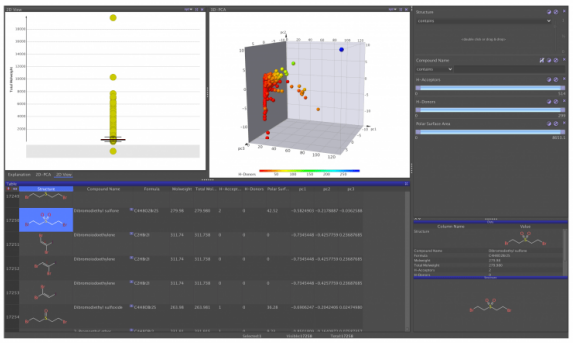

figure 7 - analyse de petites molécules chimiques avec datawarrior

auteur(s)/autrice(s) : bruno villoutreix licence : cc-by-nc

une chimiothèque est une collection de petites molécules chimiques qui peut contenir plusieurs millions de composés déjà synthétisés (ou pas). les résultats du criblage expérimental ou virtuel sont intimement liés à la qualité des composés présents dans les chimiothèques, il est donc primordial de les préparer avec soin. il existe différents types de chimiothèques, qui contiennent :

- soit des molécules issues de la synthèse chimique traditionnelle ou combinatoire,
- soit contenant des substances naturelles,
- soit des milliards de molécules qui ne sont pas encore synthétisées (il s'agit alors de collections électroniques virtuelles).

de nombreux projets actuels visent à produire des molécules virtuelles *via* des approches d'intelligence artificielle. les composés chimiques sont « entreposés » sous forme de fichiers électroniques ou de bases de données dans le cas des chimiothèques électroniques ou sous forme de poudre dans le cas de chimiothèques réelles. les fournisseurs de produits chimiques proposent de multiples collections, environ 95 millions de composés sont disponibles dans le monde. bien que ce nombre soit imposant, il ne représente qu'une infime partie des possibles. en effet, il est envisageable de synthétiser assez facilement entre 10^{20} et 10^{24} molécules mais il faut souligner que l'espace chimique total[6] est quasiment infini. ces chiffres sont impressionnants ! ils indiquent sans ambigüité qu'il est illusoire d'envisager pouvoir cribler un jour la totalité de cet espace et que les approches chémoinformatiques et de modélisation biostatistiques sont nécessaires pour construire des chimiothèques plus « intelligentes » et pour identifier des candidats médicaments originaux. on peut trouver sur internet plusieurs collections de molécules virtuelles, notamment la base de données gdb qui contient plus de 166 milliards de molécules (voir la liste disponible ci-après en pdf).

4. documents à télécharger

liste bases de donnees_drug design in silico.pdf

5. bibliographie

en plus des références mentionnées dans le texte, le lecteur est invité à consulter les ressources ci-dessous, rédigées en français.

1. vayer p, arrault a, lesur b, bertrand m, walther b. *apports de la chémoinformatique dans la recherche et l'optimisation des molécules d'intérêt thérapeutique*. med sci (paris), 2009, 25:871-7
2. bureau r. *modélisation moléculaire et conception de nouveaux ligands d'intérêts biologiques*. techniques de l'ingénieur 2014, 1-19.
3. rognan d, bonnet p. *les chimiothèques et le criblage virtuel*. med sci (paris). 2014, 30:1152-60
4. sperandio o, villoutreix b., morelli x, roche p. *les chimiothèques ciblant les interactions protéine-protéine*. med sci (paris). 2015, 31:312-9.
5. maupetit j, saladin a, tuffery p. *prédiction en ligne de la structure des protéines*. spectra analyse 2010, 276: 27-33

6. remerciements

l'auteur remercie natacha oliveira pour sa relecture attentive et ses tutos/vidéos « datawarrior ».

CRÉDITS

AUTEUR(S)/AUTRICE(S)

bruno villoutreix

bruno villoutreix est directeur de recherche à l'inserm. il travaille depuis plus de 20 ans dans le domaine de la bioinformatique structurale et de la chemoinformatique.

MISE EN LIGNE

claire vilain

responsable éditoriale de culturesciences-chimie

PARTENAIRE(S)

notes

1. **la bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. traditionnellement, la bioinformatique traite des séquences et la bioinformatique structurale de la prédiction ou de l'analyse de la structure 3d des molécules, souvent des macromolécules.**
2. **la chémoinformatique repose sur une série d'outils et de concepts qui permettent de décrire et d'exploiter des liens entre la structure et les propriétés des produits chimiques.**
3. **on peut voir que c'est une sérine protéase importante, par exemple en regardant sur pubmed, une base de données bibliographiques ouverte de référence pour effectuer des recherches dans le domaine de la santé, chimie, biologie, bioinformatique…avec plus de 30 millions de citations en janvier 2020.**
4. **ensemble de techniques pour modéliser ou simuler le comportement de molécules.**
5. **en janvier 2020, 17258 petites molécules sont présentes, certaines sont des produits naturels, d'autres des médicaments, d'autres des molécules en phases cliniques ou retirées du marché, par exemple pour des raisons de toxicité.**
6. **l'espace chimique est un concept de chemoinformatique se référant à l'espace total couvert par toutes les molécules et tous les composés chimiques qu'il est possible de synthétiser.**