PRÉDIRE LA STRUCTURE 3D DES PROTÉINES

Déterminer la structure spatiale, en trois dimensions, d'une protéine à partir de sa formule chimique est un problème aujourd'hui sans vraie réponse expérimentale. Ceci est pourtant du plus grand intérêt, en particulier pour comprendre certaines affections ou concevoir des médicaments "sur mesure". Les mécanismes d'action d'une protéine sont en effet étroitement liés à sa forme et à la façon dont elle se replie sur elle-même. La simulation numérique de ce processus a fait d'énormes progrès ces dernières années.

Il existe de nombreux problèmes pour lesquels il est très difficile, voire impossible, de trouver des réponses expérimentales. Un exemple ? La prédiction de la structure tridimensionnelle des **protéines** (voir encadré 1 dans *La modélisation des macromolécules biologiques*). Après l'élucidation du **génome** humain et de ses quelque 35 000 **gènes**, les chercheurs s'attendent à découvrir des dizaines, voire des centaines de milliers de nouvelles protéines car, contrairement à ce que l'on a longtemps pensé, un même gène peut en coder plusieurs. La fonction ainsi que les mécanismes d'action d'une protéine sont très étroitement liés à la forme géométrique de sa molécule (encadré). Or il n'existe pas encore de façon systématique de prédire la structure tridimensionnelle des protéines à partir de leur seule formule chimique. Les deux techniques de détermination de structure disponibles, la **cristallographie** et la **RMN**, ne peuvent s'appliquer à toutes les protéines et leur mise en œuvre est souvent lourde. D'autres problèmes plus conceptuels, tels les chemins et intermédiaires de repliement, ne peuvent pas être encore abordés d'un point de vue expérimental.

Pour toutes ces raisons, les scientifiques ont essayé de mettre au point des modèles et techniques de **simulation numérique** (encadré A, ***Qu'est-ce qu'une simulation numérique ?***). Ces molécules étant des assemblages d'atomes qui interagissent par des forces, il semble naturel de les **modéliser** comme un système mécanique et de faire ainsi des prédictions en résolvant les équations de base de la **mécanique de Newton**. Ce type d'approche, initié dans les années 1970, s'est considérablement développé avec la montée en puissance des ordinateurs (encadré C, ***La modélisation moléculaire***).

Une modélisation à différents niveaux

La première étape d'une simulation est la modélisation de la molécule. Suivant la résolution et le niveau de sophistication désirés, la protéine peut être modélisée comme un assemblage d'atomes liés par des ressorts ou comme un ensemble de bâtonnets qui représentent les **acides aminés**, ou même, à une échelle encore plus grossière, comme un ensemble de liens sur un réseau cubique, lorsqu'il s'agit d'étudier qualitativement certains phénomènes. Pour fixer les idées, c'est la modélisation au niveau atomique, dite *all atoms*, qui est examinée ici. La protéine y est représentée comme une collection d'atomes de carbone, oxygène, azote, hydrogène et soufre. Ces atomes sont liés entre eux par des ressorts rigides, dont la longueur correspond à celle des liaisons atomiques mesurées expérimentalement, et qui fixent la topologie de la chaîne. La molécule étant ainsi représentée, il faut aussi définir les interactions entre ses constituants. Les modèles usuels pour les **champs de forces** font intervenir des énergies d'extension pour les liens, d'élasticité pour les angles de valence(1) et des énergies de torsion pour le squelette de la chaîne. Par ailleurs, les atomes se comportant en première approximation comme des sphères dures, une interaction de type Lennard-Jones (à courte portée et très répulsive à courte distance) est introduite pour les représenter. Enfin, l'électropositivité des atomes est prise en compte par une charge partielle affectée à chacun. Ces charges partielles interagissent entre elles de façon **électrostatique**, avec une constante diélectrique(2) dont la valeur et la nature restent ardemment discutées.

Le champ de forces ou **hamiltonien** du système étant défini, la simulation numérique peut commencer. Elle se fait soit en intégrant numériquement les équations de la mécanique classique, soit en utilisant une méthode **stochastique**, de type **Monte-Carlo**, qui per-

(1) Valence : nombre de liaisons qu'un atome peut former.
(2) La constante diélectrique est la grandeur qui divise la valeur du champ électrique dans un milieu donné par rapport à ce qu'il serait dans le vide ; c'est une mesure du caractère polaire d'un solvant.

Une seule conformation parmi des milliers

Après leur synthèse par le ribosome(1), les **protéines**, molécules qui assurent la plupart des fonctions élémentaires de la cellule, se replient et acquièrent rapidement leur forme tridimensionnelle (figure 1). L'expérience a montré que les protéines se replient spontanément et que la conformation active est unique. En sorte que la structure tridimensionnelle est liée de manière univoque à la séquence primaire : parmi les milliers de conformations repliées en principe accessibles à la fibre polypeptidique(2), une seule est choisie et réalisée.
Selon quel mécanisme la protéine acquiert-elle sa conformation native ? Des progrès ont été faits au regard de cette question en étudiant expérimentalement le repliement de plusieurs molécules, dont le lysozyme. Cette molécule présente deux domaines(3) (figure 1) et il a été montré expérimentalement que, lors de son repliement, elle adopte un état transitoire dans lequel le domaine α (en **hélices**) est partiellement replié tandis que le domaine β (en **feuillets**) reste déplié. Les méthodes qui donnent ce résultat ne permettent pas d'appréhender le mécanisme au niveau atomique auquel seules des simulations de dépliement sur ordinateur permettent d'accéder.
La visualisation de l'évolution des structures secondaires pour une trajectoire de dépliement obtenues par **dynamique moléculaire** (figure 2) montre que le domaine β se déplie avant le domaine α (encadré C, ***La modélisation moléculaire***). Cette simulation indique que le repliement du domaine β est lié à la structuration de l'interface entre les deux domaines. Pour se replier, le domaine β commence par insérer deux de ses résidus (Leu56 et Ile55) dans le domaine α en cours de repliement.

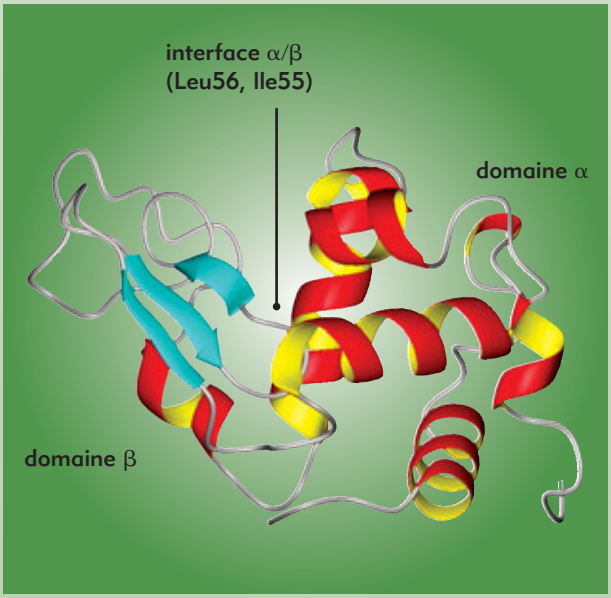


Figure 1. Représentation schématique de la structure cristalline du lysozyme. Sur la droite, le domaine α est constitué de cinq hélices (en rouge et jaune). Sur la gauche, le domaine β est constitué de trois brins β (en bleu) et d'une hélice.

(1) Le ribosome est composé d'ARN et de protéines ribosomales qui s'associent avec l'ARN messager et catalysent la synthèse des protéines.
(2) Les polypeptides sont des **polymères** linéaires composés de multiples acides aminés.
(3) Domaine : portions d'un protéine ayant une structure tertiaire (forme tridimensionnelle complexe) propre.
(4) Ces fibres sont appelées amyloïdiques par analogie avec les fibres β-amyloïdes observées dans la maladie d'Alzheimer.

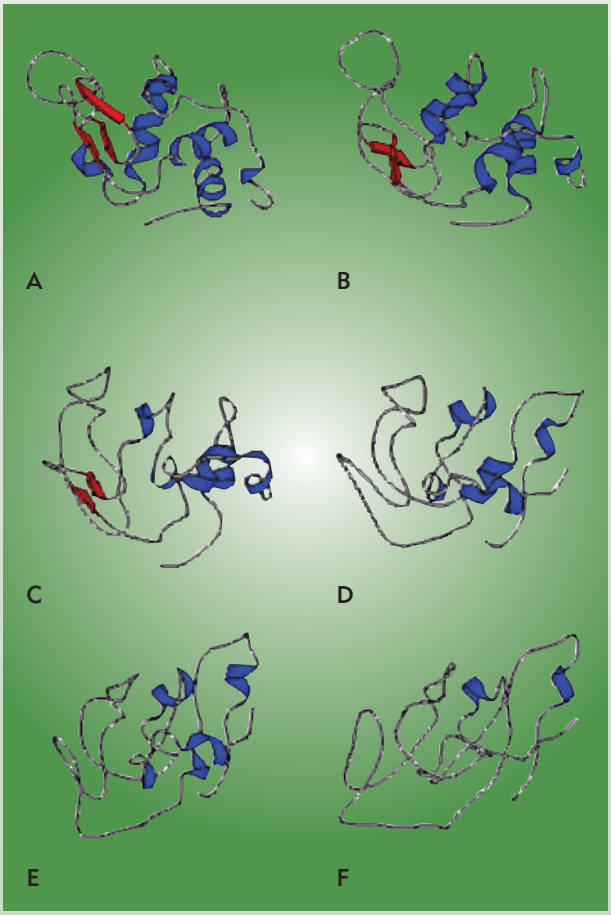


Figure 2. Représentation schématique du dépliement (de A à F) ou du repliement (de F à A) du lysozyme à six étapes différentes d'une trajectoire de dépliement obtenues par simulation de dynamique moléculaire. Les hélices α sont représentées par des serpentins en bleu et les feuillets β par des flèches rouge.

Repliements pathogènes

Ainsi, si ces résidus sont mutés d'hydrophobes en hydrophiles (en particulier, l'Ile55 muté en Thr), leur capacité à s'insérer dans le domaine α en cours de repliement est moindre et la protéine se replie moins vite. Elle va rester plus longtemps dans une forme partiellement dépliée dans laquelle le domaine β n'est pas replié. Dans cet état intermédiaire, les domaines β de deux protéines différentes peuvent s'associer et conduire à la formation d'une fibre : c'est effectivement ce qui se produit pour ce mutant du lysozyme.
D'une manière plus générale, les intermédiaires de repliement pourraient conduire à des formes mal repliées des protéines. Mal repliées ou partiellement dépliées, ces formes ont généralement tendance à s'agréger et à former des fibres amyloïdiques(4). Ainsi les intermédiaires de repliement seraient à l'origine de formes pathogènes des protéines observées dans les maladies neurodégénératives. Une protéine pourrait donc être sous sa forme correctement repliée et active ou (sous certaines conditions) adoptée sous une forme mal repliée, souvent agrégée sous forme de fibres.

Bernard Gilquin
Direction des sciences du vivant
CEA centre de Saclay

met d'échantillonner les fluctuations de la molécule dans le temps. Du fait de la présence d'interactions très répulsives à courte portée, le pas en temps utilisé pour intégrer les équations du mouvement doit être très petit, variant de la femtoseconde (10^{-15} s) à la picoseconde (10^{-12} s) suivant le degré de précision requis.

Quelques milliers de variables

La taille typique d'une protéine est d'environ 150 acides aminés, soit de l'ordre de 1 500 atomes. Chacun étant défini par trois coordonnées, il y a donc, au total, quelques milliers de variables à faire évoluer dans le temps. De plus, sachant qu'une protéine fonctionne toujours dans l'eau (il est d'ailleurs admis que l'effet hydrophobe est l'élément déterminant pour le repliement), il faut inclure dans la simulation les molécules d'eau du **solvant**, en nombre suffisant (quelques centaines dans chaque simulation) pour que l'effet hydrophobe puisse se manifester.

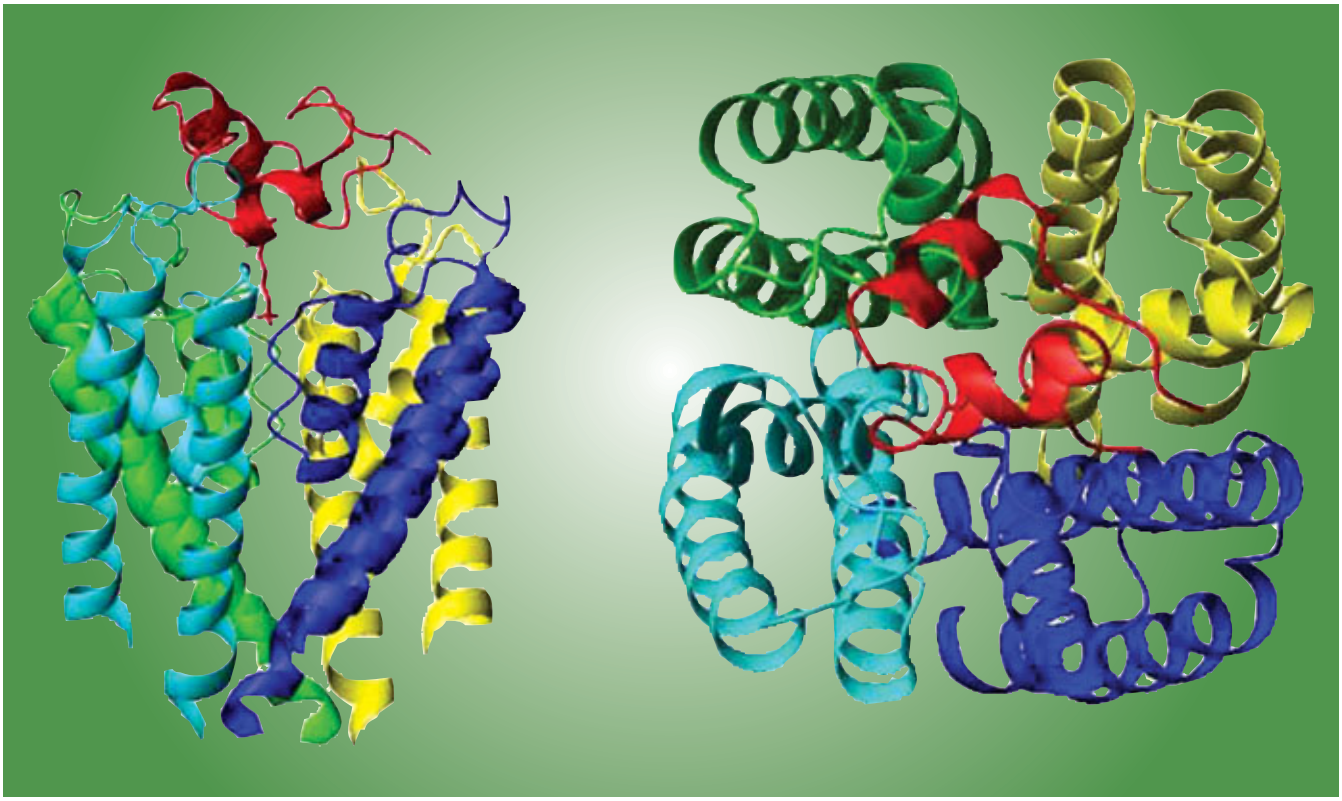
La puissance des calculateurs actuellement les plus performants permet de simuler le repliement jusqu'à des durées de l'ordre du dixième de microseconde pour les protéines les plus courtes. Les événements les plus rapides, tels la formation de structures secondaires comme les **hélices** ou les **feuillets**, apparaissent au bout de quelques microsecondes. Quant au temps de repliement lui-même, il est de l'ordre de la milliseconde. La simulation du repliement complet d'une protéine (sur plusieurs dizaines, voire centaines de millisecondes) est donc encore très lointaine, surtout si la molécule est longue.

La simulation numérique fournit cependant des informations importantes sur un certain nombre d'autres questions : détermination des modes collectifs des protéines (vibrations, grandes déformations en présence du **ligand**, etc.), chemins de repliement (quelles sont les trajectoires qui amènent la protéine de son état dénaturé à son état natif ou replié ?), raffinement des structures cristallographiques… Par ailleurs, la simulation de modèles très simplifiés sur réseau a permis de mieux comprendre la nature des séquences de protéines qui se replient et de caractériser celles dont le repliement est unique et rapide.

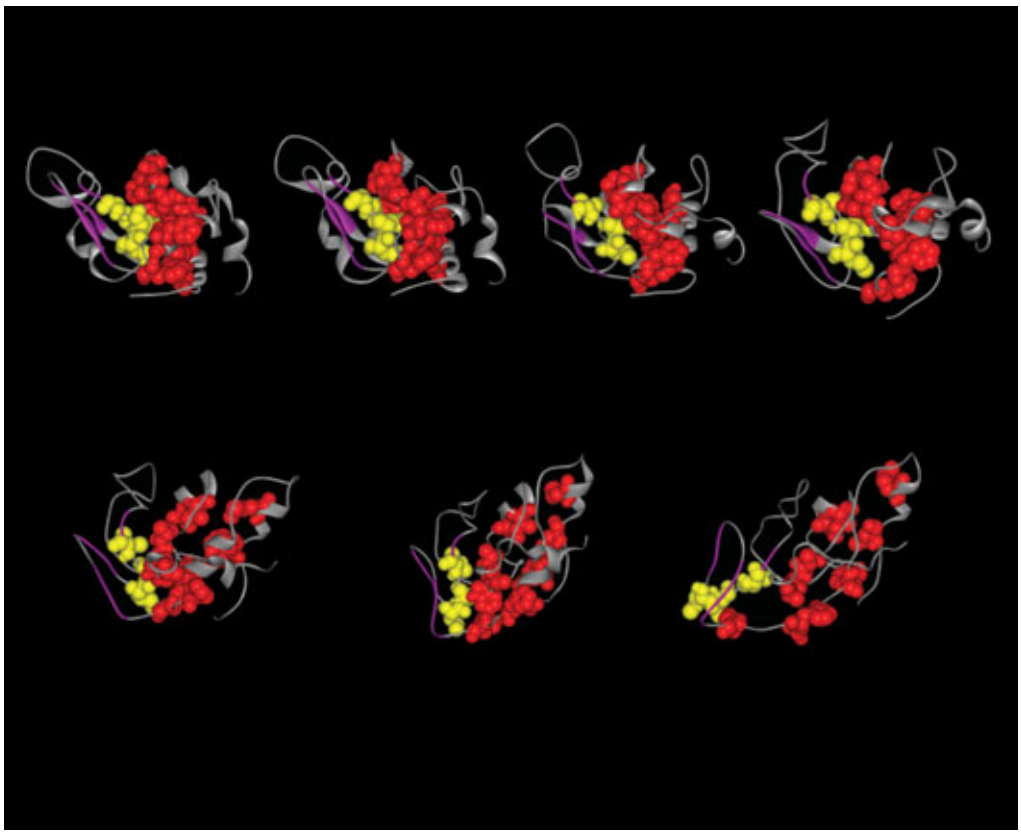
Enfin, les simulations permettent des avancées dans le domaine du *protein design*, dont le but est de synthétiser une protéine qui ait une fonction biologique bien définie ou d'améliorer sa fonction par des mutations appropriées. Elles permettent notamment d'effectuer des mutations virtuelles, ***in silico*** pourrait-on dire, sur les acides aminés du site actif : l'énergie de liaison de la protéine avec son ligand est ainsi optimisée, ce qui permet d'augmenter de façon très importante l'activité **catalytique** ou la sélectivité de la protéine. Ces prédictions peuvent ensuite être vérifiées expérimentalement.

Un problème majeur reste celui de la validité du champ de forces : ce dernier, construit de façon **semi-empirique**, est en effet ajusté sur les mesures de petites molécules ou par des **calculs quantiques** approchés (encadré C, ***La***

Modèle du complexe Bgk-Kv11, vu de face et de dessus, construit par dynamique moléculaire sous contraintes dans le cadre de l'étude des interactions de la protéine Bgk, extraite du venin d'une anémone de mer avec son récepteur le canal potassique.



LSP/CEA



J. Lunardi/CEA

Différentes phases du repliement du lysozyme, montrant l'évolution de l'interface hydrophobe entre le domaine β (en jaune) et le domaine α (en bleu). La première vue correspond à la structure de la protéine native (cf. figure 1 de l'encadré), les vues suivantes correspondent à celles de la figure 2.

modélisation moléculaire). Cependant, rien ne dit que la forme native de la protéine soit le minimum *absolu* de l'énergie du champ de forces, et donc que ce champ de forces soit à même de replier la protéine. Des chercheurs ont même montré, sur des modèles simplifiés, qu'il était impossible de trouver une paramétrisation des champs de forces telle que la forme native des protéines ait une énergie inférieure à celle de toute conformation non native.

Un concours de prédiction

Un concours de prédiction de structures en aveugle se tient tous les deux ans à Asilomar (Californie). Tous les "simulateurs" doivent prédire la structure repliée d'une protéine dont la séquence chimique, résolue expérimentalement mais non divulguée, leur est soumise comme test. La structure qu'ils proposent est ensuite comparée à la structure "secrète". Conclusion : les simulations de type *all atoms* n'ont encore aucune prédictibilité fiable au niveau des structures. Cependant, l'Américain D. Baker et son équipe de Seattle ont proposé une modélisation à une échelle intermédiaire : la protéine y est décrite comme un assemblage de fragments de trois à cinq acides aminés. Les conformations possibles de ces fragments sont échantillonnées à l'aide d'une base de données expérimentale pour constituer une bibliothèque de conformations possibles pour ces fragments. Un potentiel effectif d'interaction entre ces fragments est construit semi-empiriquement. L'interaction avec l'eau est aussi prise en compte. La simulation consiste ensuite à faire un échantillonnage **Monte-Carlo** de l'espace conformationnel, en n'admettant que les conformations tirées de la bibliothèque. Les résultats, très spectaculaires, donnent un véritable caractère prédictif à ces simulations.

Les simulations numériques ont encore d'immenses progrès à faire avant de devenir vraiment efficaces et opérationnelles, mais il est certain qu'elles constitueront prochainement un instrument prédictif indispensable pour les biologistes et les pharmacologistes.

Henri Orland
Direction des sciences de la matière
CEA centre de Saclay

Qu'est-ce qu'une simulation numérique ?

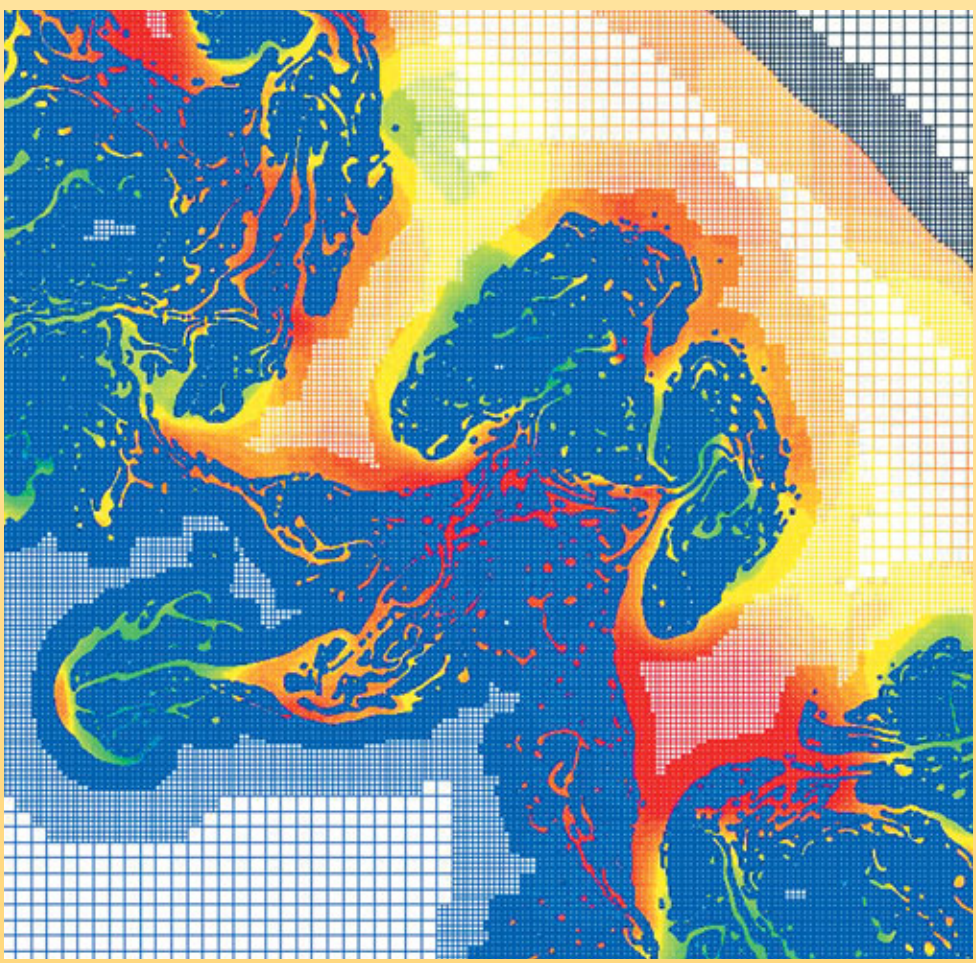
La **simulation numérique** consiste à reproduire par le calcul le fonctionnement d'un système, préalablement décrit par un ensemble de **modèles**. Elle s'appuie sur des méthodes mathématiques et informatiques spécifiques. Les principales étapes de la réalisation d'une étude par simulation numérique sont communes à de nombreux secteurs de la recherche et de l'industrie, en particulier le nucléaire, l'aérospatial ou l'automobile.

En chaque point de l'"objet" considéré, plusieurs grandeurs physiques (vitesse, température…) décrivent l'état et l'évolution du système étudié. Celles-ci ne sont pas indépendantes, mais reliées et régies par des **équations**, généralement **aux dérivées partielles**. Ces équations constituent la traduction mathématique des lois de la physique qui modélisent le comportement de l'objet. Simuler l'état de ce dernier, c'est déterminer – idéalement en tout point – les valeurs numériques de ses **paramètres**. Comme il y a un nombre infini de points, donc une infinité de valeurs à calculer, cet objectif est inaccessible (sauf dans des cas bien particuliers où l'on peut résoudre les équations de départ à l'aide de formules analytiques). Une approximation naturelle consiste donc à ne considérer qu'un nombre fini de points. Les valeurs des paramètres à calculer sont ainsi en nombre fini et les opérations nécessaires deviennent abordables grâce à l'ordinateur. Le nombre effectif de points traités dépendra bien sûr de la puissance de celui-ci : plus il sera élevé, meilleure sera finalement la description de l'objet. À la base du calcul des paramètres comme à la base de la simulation numérique, il y a donc la réduction de l'infini au fini, la **discrétisation**.

Comment opère-t-on précisément à partir des équations mathématiques du modèle ? Deux méthodes sont très souvent utilisées, respectivement représentatives des méthodes de **calcul déterministe**, qui résolvent les équations régissant les phénomènes étudiés après avoir discrétisé les variables, et des méthodes de **calcul statistique** ou **probabiliste**.

Le principe de la première, connue sous le nom de **méthode des volumes finis**, est antérieur à l'usage des ordinateurs. Chacun des points de l'objet est assimilé simplement à un petit volume élémentaire (un cube par exemple), d'où le nom de *volume fini*. Un plasma, par exemple, est ainsi vu comme un ensemble ou un réseau de volumes contigus qui, par analogie avec la trame d'un tissu, sera dénommé **maillage**. Les paramètres de l'état de l'objet sont maintenant définis dans chaque maille du maillage. Pour chacune d'elles, en reformulant les équations mathématiques du modèle par des moyennes volumiques, il sera alors possible de construire des *relations algébriques* entre les paramètres de la maille et ceux de ses voisines. Au total, il y aura autant de relations que de paramètres inconnus et ce sera à l'ordinateur de résoudre le *système* de relations obtenu. Il faudra pour cela recourir aux techniques de l'**analyse numérique** et programmer des **algorithmes** spécifiques.

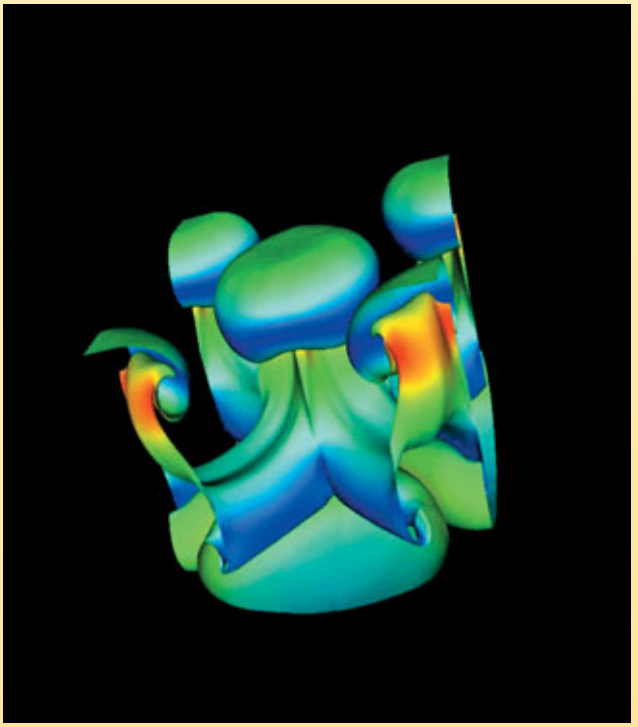
L'accroissement de la puissance des ordinateurs a permis d'augmenter la finesse de discrétisation, permettant de passer de quelques dizaines de mailles dans les années soixante à plusieurs dizaines de milliers dans les années quatre-vingt, à des millions dans les années quatre-vingt-dix et jusqu'à la dizaine de milliards de mailles aujourd'hui (machine Tera de la Direction



CEA

Exemple d'image d'une simulation 2D d'instabilités réalisée avec le supercalculateur Tera du CEA. Le calcul a fait appel au maillage adaptatif, qui se fait plus fin dans les zones où les phénomènes sont les plus complexes.

des applications militaires du CEA), chiffre qui devrait décupler à la fin de la décennie.
Un raffinement du maillage, le **remaillage adaptatif**, consiste à ajuster la taille des mailles en fonction des circonstances, par exemple en les rendant plus petites et plus serrées aux interfaces entre deux milieux, là où les phénomènes physiques sont les plus complexes, ou là où les variations sont les plus importantes.
La méthode des volumes finis s'applique dans des contextes physiques et mathématiques très variés. Elle autorise toute forme de maille (cube, hexaèdre, tétraèdre…) et le maillage peut être modifié durant le calcul, en fonction de critères géométriques ou physiques. Enfin, elle est aisée à mettre en œuvre dans le contexte des **ordinateurs parallèles** (encadré B, ***Les moyens informatiques de la simulation numérique hautes performances***), le maillage pouvant en effet faire l'objet d'un découpage pour des calculs sur ce type de machines (exemple figure B, p. 13).
Appartiennent à la même famille la **méthode des différences finies**, cas particulier de la **méthode des volumes finis** où les côtés des mailles sont orthogonaux, et la **méthode aux éléments finis**, qui peut juxtaposer divers types de mailles.
La deuxième grande méthode, dite de **Monte-Carlo**, est particulièrement adaptée pour simuler le *transport de particules*, par exemple des neutrons ou des photons d'un **plasma** (voir *Les simulations en physique des particules*). Un tel transport est en fait caractérisé par une succession d'étapes lors desquelles chaque particule peut subir différents événements (diffusion, absorption, émission…) possibles *a priori*. Les probabilités élémentaires de chacun de ces événements sont connues individuellement pour chaque particule.
Il est alors naturel d'assimiler un point du plasma à une particule. Un ensemble de particules, en nombre fini, va constituer un échantillon représentatif de l'infinité de particules du plasma, comme lors d'un sondage statistique. D'étape en étape, l'évolution de l'échantillon sera déterminée grâce à des tirages au hasard (d'où le nom de la méthode). L'efficacité de cette méthode, mise en œuvre à Los Alamos dès les années 1940, dépend bien sûr de la qualité statistique des tirages au hasard.



CEA

Simulation 3D réalisée à l'aide du supercalculateur Tera installé fin 2001 au centre CEA/DAM Île-de-France à Bruyères-le-Châtel (Essonne).

Il existe pour cela des méthodes de nombres aléatoires, bien adaptées au traitement par un ordinateur.
Les méthodes des volumes finis et de Monte-Carlo ont suscité et suscitent de nombreuses études mathématiques. Ces études s'attachent notamment à préciser la convergence de ces méthodes, c'est-à-dire comment la précision de l'approximation varie avec le nombre de mailles ou de particules. Cette question est naturelle lors de la confrontation des résultats de la simulation numérique à ceux de l'expérience.

Comment se déroule une simulation numérique ?

Il est souvent question d'*expérience numérique* pour souligner l'analogie entre la pratique d'une simulation numérique et la conduite d'une expérience de physique.
Brièvement, cette dernière utilise un dispositif expérimental, configuré selon des conditions initiales (de température, de pression…) et des paramètres de contrôle (durée de l'expérience, des mesures…). Durant l'expérience, le dispositif produit des points de mesures qui sont enregistrés. Ces enregistrements sont ensuite analysés et interprétés.
Dans une simulation numérique, le dispositif expérimental consiste en un ensemble de programmes informatiques exécutés sur des ordinateurs. Les **codes** ou **logiciels de calcul** sont la traduction, à travers des algorithmes numériques, des formulations mathématiques des modèles physiques étudiés. En amont et en aval du calcul, les *logiciels d'environnement* effectuent la gestion de plusieurs opérations complexes de préparation des calculs et de leur dépouillement.
Les données initiales de la simulation comporteront d'abord la délimitation du domaine de calcul à partir d'une représentation approchée des formes géométriques (produite par le dessin et la CAO, conception assistée par ordinateur), suivie de la discrétisation de ce domaine de calcul sur un maillage, ainsi que les valeurs des paramètres physiques sur ce maillage et les paramètres de contrôle du bon déroulement des programmes…Toutes ces données (produites et gérées par les logiciels d'environnement) seront saisies et vérifiées par les codes. Les résultats des calculs proprement dits, c'est-à-dire les valeurs numériques des paramètres physiques, seront sauvegardés au fur et à mesure. En fait, un protocole spécifique structurera les informations produites par l'ordinateur afin de constituer une base de données numériques.
Un protocole complet organise l'échange informatique des informations requises (dimensions notamment) suivant des formats prédéfinis : modeleur(1), mailleur(2), découpeur de maillage, code

(1) Le modeleur est un outil qui permet la création et la manipulation de points, courbes et surfaces en vue par exemple de la création d'un maillage.
(2) Les formes géométriques d'un maillage sont décrites par des ensembles de points reliés par des courbes et des surfaces (de Bézier par exemple) qui en représentent les frontières.

de calculs, logiciel de visualisation et d'analyse. Les études de *sensibilité* des résultats (au maillage et aux modèles) font partie des "expériences" numériques.
À l'issue des calculs (résolution numérique des équations décrivant les phénomènes physiques qui se déroulent dans chaque maille), l'analyse des résultats par des spécialistes reposera sur l'exploitation de la base de données numériques. Elle comportera plusieurs étapes : extraction sélective des données (selon le paramètre physique recherché) et visualisation, extraction et transfert des données pour calculer et visualiser des diagnostics.
Le parallèle entre la conduite d'un cas de calcul, d'une expérience numérique et la conduite d'une expérience physique ne s'arrête pas là : les résultats numériques seront comparés aux résultats expérimentaux. Cette analyse comparative, effectuée sur la base de critères quantitatifs standardisés, fera appel et à l'expérience et à l'art de l'ingénieur, du physicien, du mathématicien. Elle débouchera sur de nouvelles améliorations des modèles physiques et des programmes informatiques de simulation.

Bruno Scheurer
Direction des applications militaires
CEA centre DAM-Ile de France

Frédéric Ducros et **Ulrich Bieder**
Direction de l'énergie nucléaire
CEA centre de Grenoble

L'exemple d'un calcul de thermohydraulique

La mise en œuvre d'un protocole de simulation numérique peut être illustrée par les travaux réalisés par l'équipe de développement du logiciel de calcul **thermohydraulique** Trio U. Ces travaux se sont déroulés dans le cadre d'une étude faite en collaboration avec l'Institut de radioprotection et de sûreté nucléaire (IRSN). L'objectif était d'obtenir des données très précises pour fournir à l'ingénieur les sollicitations en température à la paroi des composants d'un réacteur à eau sous pression dans le cas d'un accident grave impliquant une circulation naturelle turbulente de gaz chauds. Cette étude requiert la modélisation simultanée d'effets "système" à grande échelle et de phénomènes **turbulents** à petite échelle (encadré F, ***Modélisation et simulation des écoulements turbulents***).
Elle débute par la définition du modèle de calcul global (figure A), suivie de la réalisation du modèle CAO et du maillage correspondant avec des logiciels du commerce (figure B). Les maillages de plus de cinq millions de mailles exigent l'utilisation de puissantes stations graphiques. Dans cet exemple, le maillage d'un générateur de vapeur (figures C et D) a été découpé pour répartir les calculs sur huit processeurs d'un calculateur parallèle du CEA : chaque couleur symbolise une zone affectée à un processeur particulier. Les calculs, dont les conditions aux limites sont données par un calcul "système" (Icare-Cathare), produisent des résultats qu'il appartient aux spécialistes d'interpréter. En l'occurrence, les visualisations sur des stations graphiques des valeurs instantanées des champs de vitesse montrent l'impact d'un panache chaud sur la plaque tubulaire du générateur de vapeur (coupe dans le champ de vitesses à gauche de la figure E) et la température instantanée dans la boîte à eau (à droite).

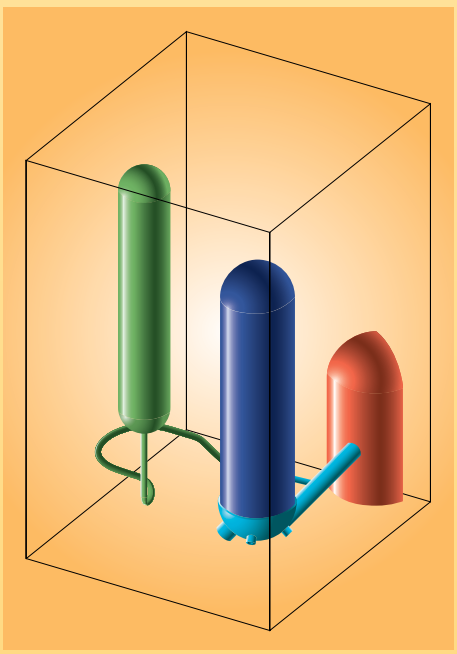


***Figure A.** Domaine de calcul global incluant une partie de la cuve réacteur (rouge), la conduite de sortie (branche chaude en bleu clair), le générateur de vapeur (bleu foncé) et le pressuriseur (vert).*

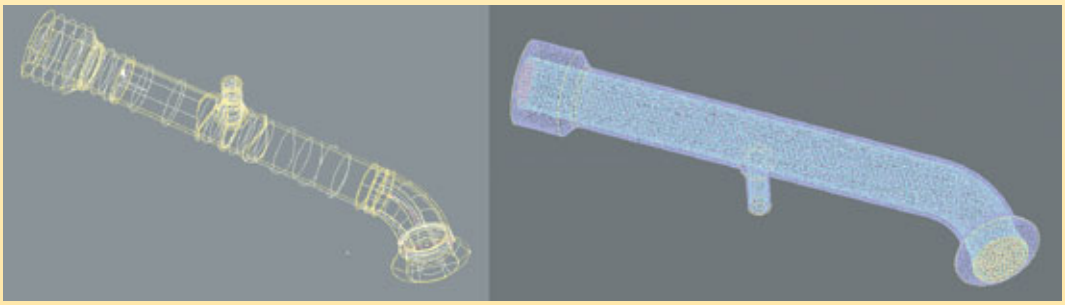
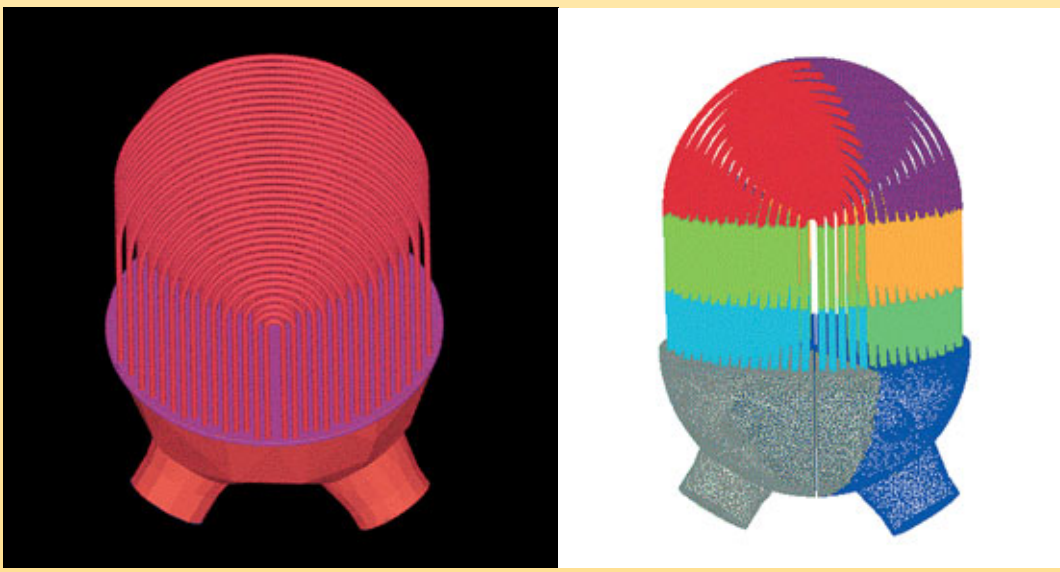


Figure B. Modèle CAO de la branche chaude en sortie de la cuve réacteur (à gauche) et son maillage non structuré (à droite).



Figures C et D.

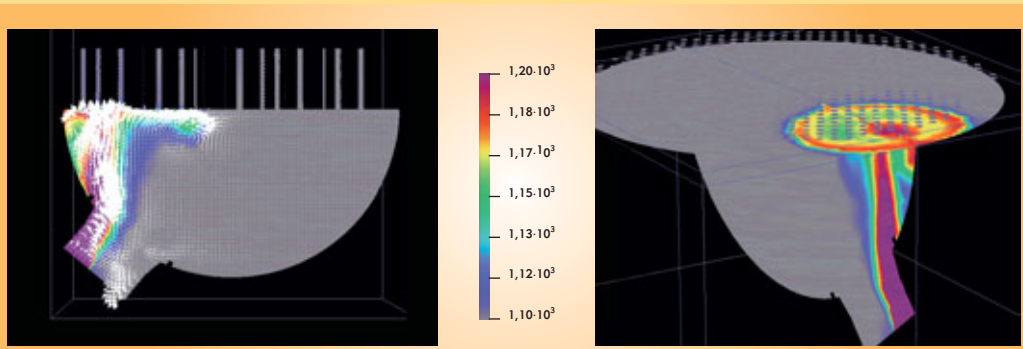


Figure E.

Les moyens informatiques de la simulation

Effectuer des **simulations numériques** plus précises impose de mettre en œuvre des **modèles** physiques et numériques eux-mêmes plus précis portant sur des descriptions plus fines des objets simulés (encadré A, ***Qu'est-ce qu'une simulation numérique ?***). Tout ceci nécessite des progrès dans le domaine des logiciels de simulation mais aussi une augmentation importante de la capacité des équipements informatiques sur lesquels ces logiciels sont utilisés.

Processeurs scalaires et vectoriels

Au cœur de l'ordinateur, le processeur est l'unité de base qui, déroulant un programme, effectue les calculs. Il en existe deux grands types, les **processeurs scalaires** et les **processeurs vectoriels**. Les premiers exécutent des opérations portant sur des nombres élémentaires (scalaires), par exemple l'addition de deux nombres. Les seconds exécutent des opérations portant sur des ensembles de nombres (vecteurs), par exemple additionner deux à deux les nombres composant deux ensembles de 500 éléments. À ce titre, ils sont particulièrement adaptés à la simulation numérique : lors de l'exécution d'une opération de ce type, un processeur vectoriel peut fonctionner à une vitesse proche de sa performance maximale (crête). La même opération avec un processeur scalaire exige de nombreuses opérations indépendantes (opérations par composante des vecteurs) qui s'exécutent à une vitesse bien inférieure à sa vitesse crête. L'avantage principal des processeurs scalaires est leur prix : il s'agit de microprocesseurs généralistes dont les coûts de conception et de fabrication peuvent être amortis sur de larges marchés.

Forces et contraintes du parallélisme

Les processeurs récents permettent de hautes performances, d'une part en utilisant une fréquence de fonctionnement plus élevée, d'autre part en cherchant à exécuter en même temps plusieurs opérations : c'est un premier niveau de **parallélisme**. L'accélération de la fréquence est limitée par l'évolution de la technologie micro-électronique, tandis que les dépendances entre instructions à exécuter par le processeur limitent le parallélisme possible. La mise en œuvre simultanée de plusieurs processeurs constitue un second niveau de parallélisme, qui permet d'obtenir des performances accrues à condition de disposer de programmes capables d'en tirer parti. Alors que le parallélisme au niveau des processeurs est automatique, celui *entre processeurs* dans un ordinateur parallèle est à la charge du programmeur, qui doit découper son programme en morceaux indépendants et prévoir entre eux les communications nécessaires. On procède souvent par un découpage du domaine sur lequel porte le calcul, chaque processeur étant chargé de simuler le comportement d'un domaine, et par l'établissement de communications régulières entre processeurs afin de garantir la cohérence d'ensemble du calcul. Pour obtenir un programme parallèle efficace, il faut s'assurer de l'équilibrage de charge entre processeurs et chercher à limiter le coût des communications.

Les différentes architectures

Les équipements informatiques ont différentes fonctions. À partir de son ordinateur de travail sur lequel il prépare ses calculs et en analyse les résultats, l'utilisateur accède à des moyens de calcul, de stockage, et de visualisation partagés, mais beaucoup plus puissants que les siens propres. L'ensemble de ces équipements sont reliés par des réseaux informatiques permettant de faire circuler les informations entre eux avec des débits compatibles avec le volume de données produites, pouvant atteindre 1 **téraoctet** (1 To = 10^{12} octets) de données pour une seule simulation.

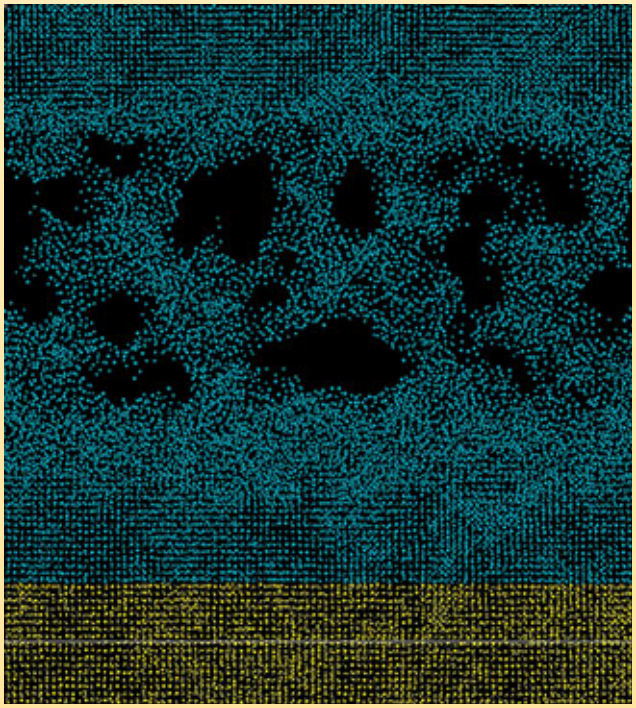
Les grands équipements de calcul sont généralement appelés **supercalculateurs**. Ils atteignent aujourd'hui des puissances qui se chiffrent en **téraflops** (1 Tflops = 10^{12} opérations de calcul par seconde).

Il existe aujourd'hui trois grands types de supercalculateurs : les supercalculateurs vectoriels, les grappes de mini-ordinateurs à mémoire partagée et les grappes de PC (l'ordinateur que chacun possède chez soi). Le choix entre ces architectures dépend largement des applications et de l'utilisation visées. Les supercalculateurs vectoriels disposent de processeurs très performants mais dont il est difficile d'augmenter la puissance en ajoutant des processeurs. Les grappes de PC sont peu coûteuses mais mal adaptées à des environnements où de nombreux utilisateurs font beaucoup de calculs très gourmands en puissance machine, en mémoire et en entrées-sorties.

Ce sont ces considérations qui ont en particulier conduit la Direction des applications militaires (DAM) du CEA à choisir pour son programme simulation (voir *Le programme Simulation : la garantie des armes sans essais nucléaires*) les architectures de type grappe de mini-ordinateurs à mémoire partagée, encore appelées ***clusters* de SMP** (Symmetric Multi-Processor). Un tel système utilise comme brique de base un mini-ordinateur com-

CEA

Installée en décembre 2001 au CEA (centre DAM-Ile de France) et conçue par Compaq (devenue depuis HP), la machine Tera a pour élément de base un mini-ordinateur à 4 processeurs Alpha à 1 GHz partageant une mémoire de 4 Go et fournissant une puissance totale de 8 Gflops. Ces éléments de base sont interconnectés par un réseau rapide conçu par la société Quadrics. Une opération de synchronisation sur l'ensemble des 2 560 processeurs s'effectue en moins de 25 microsecondes. Le système de fichiers global offre un espace de stockage de 50 téraoctets pour les entrées-sorties avec une bande passante agrégée de 7,5 Go/s.



CEA

Les calculateurs parallèles sont adaptés aux méthodes numériques basées sur des maillages (encadré A, Qu'est-ce qu'une simulation numérique ?) mais aussi au traitement de calculs ab initio *comme cette simulation par dynamique moléculaire de l'endommagement par choc de deux plaques de cuivre à 1 km/s (voir* La simulation des matériaux*). Le système considéré est constitué de 100 000 atomes de cuivre représentant un parallélépipède de section carrée (0,02 µm de côté) à densité normale. Les atomes interagissent suivant un potentiel EAM (*embedded atom potential*) pendant 4,6 picosecondes. Le calcul, effectué sur 18 processeurs du supercalculateur Tera de Bruyères-le-Châtel à l'aide du logiciel Stamp développé au CEA, a représenté une dizaine de minutes de temps "utilisateur" (calcul réalisé par B. Magne). Des tests impliquant jusqu'à 64 millions d'atomes ont été réalisés, mobilisant 256 processeurs pendant une centaine d'heures.*

portant plusieurs microprocesseurs qui partagent une mémoire commune (figure). Ces mini-ordinateurs étant largement diffusés dans des domaines variés allant de la banque au serveur web en passant par les bureaux d'études, ils offrent un excellent rapport performance/prix. Ces "briques" de base (encore appelées *nœuds*) sont reliées entre elles par un réseau d'interconnexion hautes performances : la puissance cumulée de plusieurs centaines de ces "briques" peut atteindre plusieurs téraflops. On parle alors d'**ordinateur massivement parallèle**.

Cette puissance peut être disponible pour une seule application parallèle utilisant toutes les ressources du supercalculateur mais aussi pour de multiples applications indépendantes, parallèles ou non, utilisant chacune une partie des ressources.

Si la caractéristique mise en avant pour décrire un supercalculateur est en général sa puissance de calcul, il ne faut pas négliger l'aspect entrées-sorties. Ces machines capables d'effectuer des simulations de grande taille doivent disposer de systèmes de disques avec des capacités et des performances adaptées. Dans les *clusters* de SMP, chaque mini-ordinateur dispose d'un espace disque local. Il n'est néanmoins pas judicieux d'utiliser celui-ci pour les fichiers utilisateurs, ce qui obligerait l'utilisateur à explicitement déplacer ses données entre les différentes phases de ses calculs. Pour cette raison, il est important de disposer d'un espace disque accessible par l'ensemble des mini-ordinateurs du supercalculateur. Cet espace est en général constitué de batteries de disques reliées à des nœuds dont la fonction principale est de les gérer. Comme pour le calcul, c'est le parallélisme pour les entrées-sorties qui permet d'offrir des performances élevées. Il faut, pour ce faire, disposer de systèmes de fichiers globaux parallèles permettant un accès rapide et sans contraintes à l'espace disque partagé.

Offrant des puissances de calcul considérables, les *clusters* de SMP posent néanmoins plusieurs défis. Parmi les plus importants, outre la programmation de logiciels de simulation capables de tirer parti du grand nombre de processeurs, il faut mettre au point des systèmes d'exploitation et les logiciels associés compatibles avec de telles configurations et tolérants vis-à-vis des pannes.

François Robin
Direction des applications militaires
CEA centre DAM-Ile de France

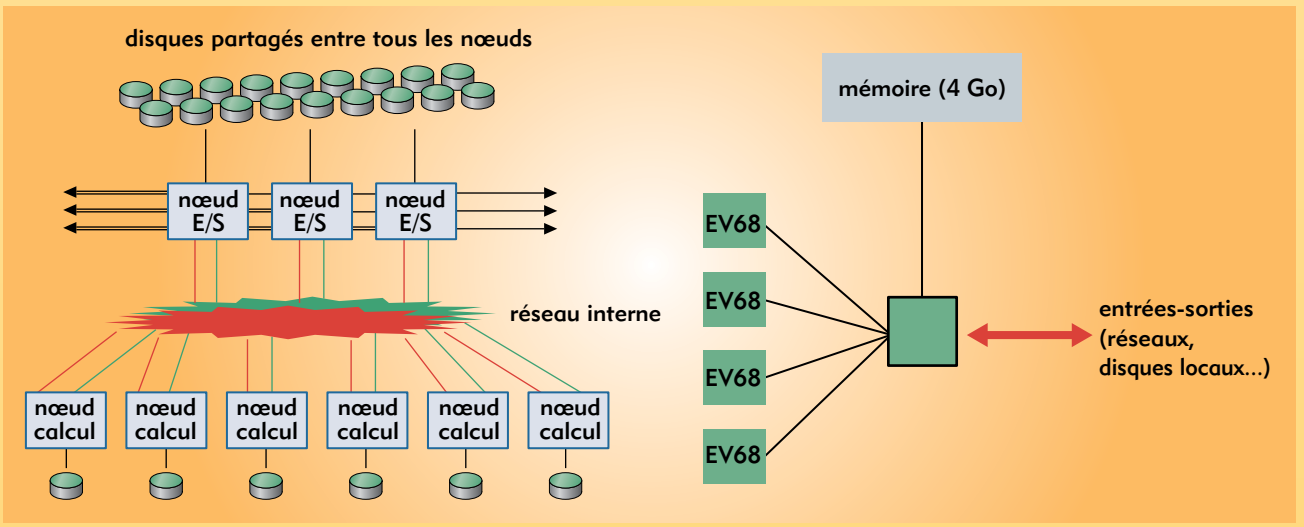


*Figure. Architecture d'une machine du type "*cluster *de SMP". À gauche, l'architecture générale (E/S = entrée/sortie), à droite celle d'un nœud avec quatre processeurs Alpha EV68 cadencés à 1 GHz.*

Modélisation et simulation des écoulements turbulents

La **turbulence**, ou l'agitation de l'écoulement dit turbulent, se développe dans la plupart des écoulements qui conditionnent notre environnement immédiat (rivières, océan, atmosphère). Elle se révèle être aussi un, sinon le, paramètre dimensionnant dans un bon nombre d'écoulements industriels (liés à la production ou la conversion d'énergie, à l'aérodynamique…). Il n'est donc pas étonnant que soient entrepris des efforts visant sa prédiction – fût-elle encore imprécise – surtout lorsqu'elle se trouve combinée à des phénomènes qui la compliquent : stratification, combustion, présence de plusieurs phases… C'est que, paradoxalement, même s'il est possible d'anticiper la nature turbulente d'un écoulement et même, d'un point de vue théorique, de dégager certaines caractéristiques communes et apparemment universelles aux écoulements turbulents[1], leur prédiction dans des cas précis reste délicate. Celle-ci doit en effet prendre en compte l'importante gamme d'échelles spatiales et temporelles[2] impliquées dans tout écoulement de ce type.

Les chercheurs ne sont pourtant pas démunis, aujourd'hui, pour aborder ce problème. En premier lieu, les équations qui régissent l'évolution spatio-temporelle des écoulements turbulents (équations de Navier-Stokes[3]) sont connues. Leur résolution complète, dans des cas très favorables, a conduit à des descriptions prédictives. Mais l'emploi systématique de cette méthode de résolution se heurte à deux difficultés rédhibitoires : d'une part, il nécessiterait la connaissance complète et simultanée de toutes les variables attachées à l'écoulement et des forçages s'exerçant sur lui[4] et, d'autre part, il mobiliserait des moyens de calculs irréalistes pour encore des décennies.

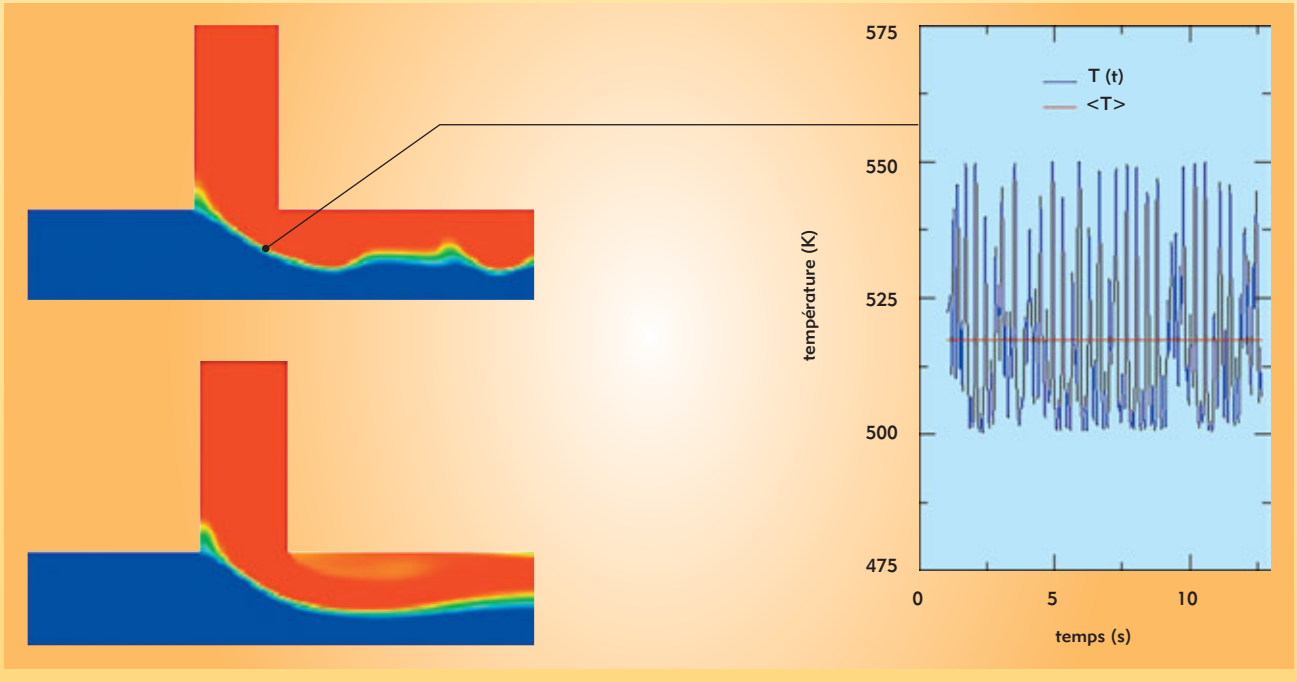


Figure. Champ de température instantané (haut) et moyenné (bas) dans une situation de mélange. La courbe donne l'historique de la température en un point : valeur instantanée fluctuante en bleu et moyenne en rouge (d'après la thèse d'Alexandre Chatelain [DEN/DTP/SMTH/LDTA]).

Il faut donc se résoudre, en s'appuyant sur le caractère fluctuant dû à l'agitation turbulente, à définir et utiliser des moyennes. Une des approches les plus répandues consiste à aborder le problème sous un angle statistique. Les moyennes d'ensemble de vitesse, de pression, de température… dont la distribution caractérise l'écoulement turbulent sont définies comme les variables principales de l'écoulement qu'on cherche à qualifier par rapport à ces moyennes. Ceci conduit à une décomposition du mouvement (dite de Reynolds) en champs moyen et fluctuant, ce dernier mesurant l'écart instantané et local entre chaque grandeur réelle et sa moyenne (figure). Ces fluctuations représentent la turbulence et couvrent une partie importante du spectre de Kolmogorov(1).

Cette opération réduit considérablement le nombre de degrés de liberté du problème et le rend « manipulable » informatiquement. Elle comporte aussi de nombreuses difficultés : il faut tout d'abord constater que, précisément en raison des non-linéarités des équations du mouvement, toute moyenne fait surgir des termes nouveaux et inconnus qu'il faut estimer. En fermant la porte à la description complète et déterministe du phénomène, on ouvre celle de la modélisation, c'est-à-dire à la représentation des effets de la turbulence sur les variables moyennes.

Beaucoup de progrès ont été accomplis depuis les premiers modèles (Prandtl, 1925). Les modélisations n'ont cessé d'évoluer vers plus de complexité, se basant sur le fait généralement vérifié que toute nouvelle extension permet de conserver les propriétés antérieurement acquises. Il faut aussi constater que, même si de nombreux développements remettent en avant la nécessité de traiter les écoulements en respectant leur caractère *instationnaire*, les modélisations les plus populaires ont été développées dans le cadre des écoulements *stationnaires*, pour lesquels on n'accède donc qu'à une représentation de la moyenne temporelle de l'écoulement : dans le modèle mathématique final, les effets de la turbulence proviennent ainsi intégralement de la modélisation.

Il est également remarquable que, malgré de nombreux travaux, aucune modélisation n'est aujourd'hui capable de rendre compte de l'intégralité des phénomènes qui influencent la turbulence ou sont influencés par elle (transition, instationnarité, stratification, compression, etc.). Ce qui semble pour l'instant empêcher les modélisations statistiques de nourrir une ambition d'universalité.

Malgré ces limitations, la plupart des modélisations statistiques courantes sont maintenant disponibles dans les codes commerciaux et les outils des industriels. Il n'est pas possible de prétendre qu'elles permettent des calculs prédictifs dans toute situation. Leur précision est variable, offrant des résultats utiles pour l'ingénieur dans des situations maîtrisées et favorables (prédiction de la trainée avec une précision de 5 % à 10 % d'erreur [parfois mieux] sur certains profils), mais parfois faux dans des situations qui se révèlent, après coup, en dehors du champ de validité du modèle. Tout emploi maîtrisé d'une modélisation repose donc sur une qualification particulière au type d'écoulement à traiter.

Des modélisations alternatives, répondant au besoin d'une plus grande précision sur des gammes d'échelles spatiales et temporelles plus étendues et donc basées sur un opérateur de "moyenne" d'une nature différente, sont actuellement en développement et représentent des voies nouvelles.

Le paysage des modélisations de la turbulence est aujourd'hui très complexe et l'unification des points de vue et des divers concepts de modélisation est une gageure. La tentation de l'universalité des modélisations reste donc hors de propos. Leur mise en œuvre réelle relève la plupart du temps de compromis généralement guidés par le savoir-faire de l'ingénieur.

Frédéric Ducros
Direction de l'énergie nucléaire
CEA centre de Grenoble

(1) On peut faire référence à la répartition spectrale de l'énergie cinétique turbulente, connue comme le "spectre de Kolmogorov", qui illustre de manière très simple la hiérarchie des échelles, des grandes échelles porteuses d'énergie aux échelles de plus en plus petites et de moins en moins énergétiques.

(2) Cette étendue est le résultat des non-linéarités des équations du mouvement qui donne naissance à une gamme étendue d'échelles spatiales et temporelles. Cette gamme est une fonction croissante du nombre de Reynolds, Re, mesurant le rapport entre force d'inertie et force visqueuse.

(3) L'hypothèse selon laquelle la résolution complète des équations de Navier-Stokes permet la simulation de la turbulence est généralement admise, tout du moins dans la gamme des écoulements sans choc.

(4) Il s'agit d'un problème régi par des conditions initiales et aux limites.

La modélisation moléculaire

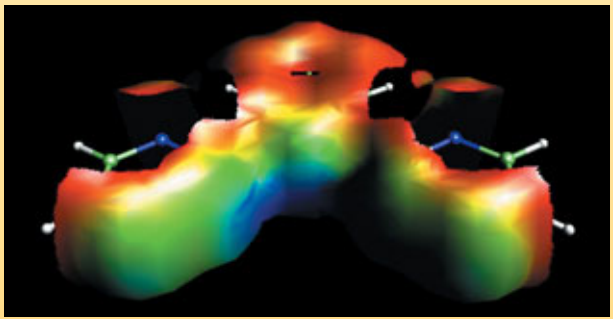
Les chercheurs en chimie et en biologie, ainsi qu'en physique des matériaux, font de plus en plus largement appel à des outils qui leur permettent de prévoir avec toujours plus de finesse les effets des molécules en fonction de leur structure et, ce qui est encore plus intéressant pour les concepteurs, d'imaginer la structure de nouveaux édifices qui est susceptible d'induire un comportement donné et *in fine* produire l'effet recherché.

Cette approche **théorique** dispose d'une large gamme d'outils mais parmi les principaux figurent la chimie quantique, la mécanique et la dynamique moléculaires, également utilisées en physique des matériaux.

La **chimie quantique**, basée sur les lois de la mécanique quantique, sert avant tout à décrire la structure des molécules et leur comportement dans les processus tels que les réactions chimiques.

La **dynamique moléculaire** *classique* simule le mouvement des atomes des systèmes moléculaires et l'évolution de leur configuration spatiale à partir d'équations de la mécanique classique. Elle donne accès à des propriétés structurales, dynamiques et thermodynamiques.

Comme la chimie quantique, la **mécanique moléculaire** est une méthode permettant d'étudier la structure et le comportement des molécules mais elle est moins coûteuse, plus rapide et, donc, peut être utilisée pour décrire les systèmes de milliers d'atomes tels que les **macromolécules biologiques**.



CEA/DEN/J.-P. Dognon

Représentation, calculée par une méthode de chimie quantique, du potentiel électrostatique autour d'une molécule de BTP (bis-triazinyl-pyridine) développée pour le procédé Sanex de séparation des actinides et des lanthanides.

Les macromolécules biologiques

Les cellules sont les entités fondamentales de tous les organismes vivants. En dehors de l'eau, les **macromolécules** biologiques sont les constituants majeurs de la cellule, où elles remplissent des fonctions multiples. Une macromolécule biologique est composée de sous-unités de faible poids moléculaire, ajoutées les unes aux autres pour donner un long **polymère** en forme de chaîne. Habituellement, chaque chaîne n'est formée que d'une seule famille de sous-unités dont l'enchaînement précis est essentiel à la fonction de la macromolécule. Les grandes catégories de macromolécules sont au nombre de quatre.

Les protéines sont probablement les macromolécules les plus importantes parce qu'elles jouent un rôle prédominant dans la plupart des processus biologiques. Par exemple, les **enzymes** sont les protéines qui **catalysent** la majorité des réactions chimiques dans la cellule. D'autres classes de protéines ont plutôt un rôle structural ou sont impliquées dans la signalisation(1), la régulation du **métabolisme** ou la défense immunitaire. Les protéines sont des polymères d'**acides aminés** – dont une vingtaine de types différents sont communément trouvés – et une protéine peut être constituée de quelques chaînes, chacune comportant quelques centaines d'acides aminés. Les protéines sont souvent associées à d'autres molécules qui les assistent dans leurs tâches biologiques. Les structures tridimensionnelles des protéines sont très compliquées mais critiques pour leur fonctionnement.

Les acides nucléiques – l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN) – sont des polymères de **nucléotides**. L'**ADN**, dans sa forme double brin (deux chaînes de nucléotides organisées en double hélice), est le matériel génétique qui, entre autre, code les instructions pour les séquences (l'enchaînement des acides aminés) de toutes les protéines synthétisées par la cellule. L'**ARN**, habituellement dans une forme simple brin (une chaîne de nucléotides), est essentiel pour la synthèse des protéines.

Les lipides, constituants fondamentaux des membranes cellulaires, exercent également une fonction importante dans le métabolisme et comme réservoirs énergétiques. Parmi les classes de lipides se trouvent les phospholipides, les triglycérides et les stéroïdes.

Les polysaccharides sont des polymères de sucres simples, tels que le fructose ou le glucose. Ils jouent un rôle structural, notamment dans les plantes (la cellulose est un polysaccharide), dans la reconnaissance moléculaire, et peuvent servir de réservoirs énergétiques.

(1) Transmission de signaux permettant aux cellules de communiquer entre elles.