IFT 3290 Hiver 2006

Bioinformatique

Miklós Csűrös

André-Aisenstadt 3149
csuros@iro.umontreal.ca

http ://www.iro.umontreal.ca/~csuros/IFT3290/

Plan de cours

Préalables :

IFT2010 — Programmation 2

IFT1978 — Probabilités et statistique

Horaire :

cours	mercredi 11 :30–13 :30	Z 215
cours	vendredi 14 :30–16 :30	Z 350

Évaluation : 30% TP, 30% intra, 30% final, 10% présentation

4 travaux : programmation et théorie

Diversité et universalité

Diversité de la vie

Universalité au niveau moléculaire : interaction de macromolécules, mécanismes communes (évidence de l'Évolution)

Protéines et acides nucléiques

Protéines

Fonctions :

- structure de la cellule [p.e collagène form des fibres en tissu conjonctif]
- enzymes : catalysateur de réactions spécifiques
- protéines membranaires
- signalisation [p.e. EGF — facteur de croissance pour l'épiderme]
- ...

EXEMPLE : KINESINE

[kinesine]

(moteur moléculaire)

Info sur une protéine

facteur de croissance EGF : voie de signalisation (www)

recherche dans une base de donnée : SWISS-PROT (www)

(Tapez «EGF receptor» au site Web de SWISS-PROT et suivez le lien vers EGFR_HUMAN.)

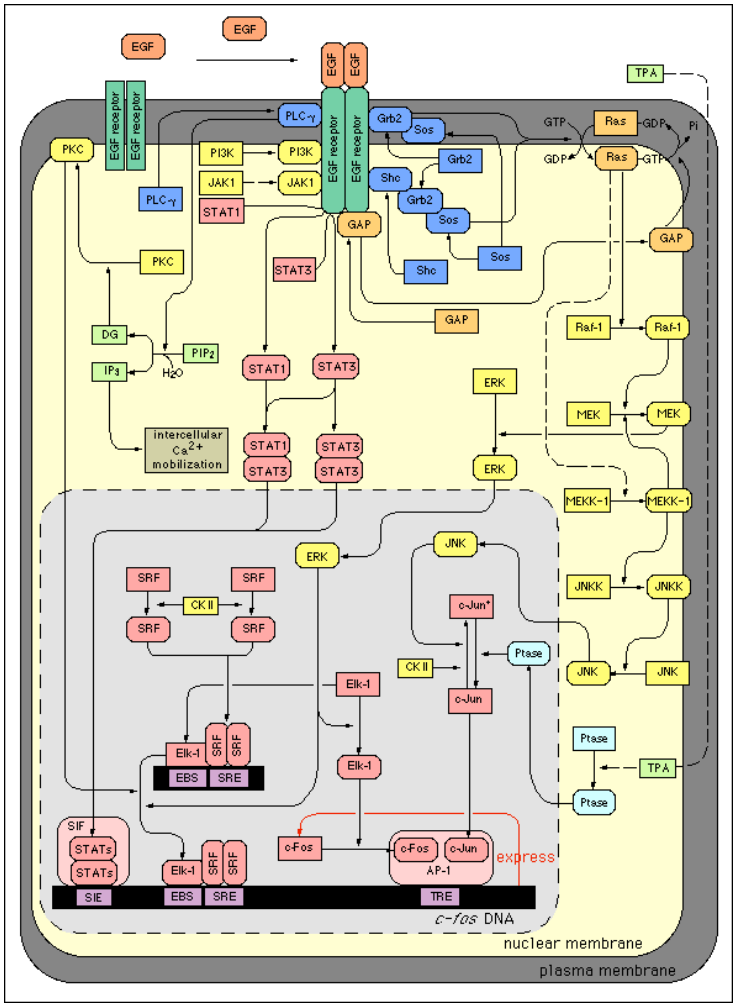
pleine d'information liée : e.g., structure 3D (www)

(Suivez le lien «PDB» dans la ligne «HSSP».)

Séquence

```
MRPSGTAGAA LLALLAALCP ASRALEEKKV CQGTSNKLTQ
LGTFEDHFLS LQRMFNNCEV VLGNLEITYV QRNYDLSFLK
...
```

VOIE DE SIGNALISATION EGF



SPAD, Kyushu University

EGF DANS SWISS-PROT

NicePro View of SWISS-PROT: P00533 1/10/03 9:42 AM

ExPASy Home page | Site Map | Search ExPASy | Contact us | SWISS-PROT

Hosted by CBR Canada Mirror sites: Bolivia China Korea Switzerland Taiwan USA

NicePro View of SWISS-PROT: P00533

Printer-friendly view | Quick BlastP search

[General] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

nombre d'accession

nom de la BD

General information about the entry

Entry name	**EGFR_HUMAN**
Primary accession number	**P00533**
Secondary accession numbers	P06268 Q14225 Q9UMD7 ... O00688 Q9BZS2 Q9H2C9 Q9GZX1 Q9H3C9
Entered in SWISS-PROT in	Release 01, July 1986
Sequence was last modified in	Release 35, November 1997
Annotations were last modified in	Release 41, June 2002

Name and origin of the protein

Protein name	**Epidermal growth factor receptor [Precursor]**
Synonyms	**EC** 2.7.1.112 **Receptor protein-tyrosine kinase ErbB-1**
Gene name	**EGFR** or **ERBB1**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] SEQUENCE FROM NUCLEIC ACID (ISOFORM 1).
MEDLINE=84219729; PubMed=6328312; [NCBI, ExPASy, EBI, Israel, Japan]
Ullrich A., Coussens L., Hayflick J.S., Dull T.J., Gray A., Tam A.W., Lee J., Yarden Y., Libermann T.A., Schlessinger J., Downward J., Mayes E.L.V., Whittle N., Waterfield M.D., Seeburg P.H.;
"Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells.";
Nature 309:418-425(1984).

[2] SEQUENCE FROM NUCLEIC ACID (ISOFORM 2).
TISSUE=Placenta;
MEDLINE=95382957; PubMed=7654368; [NCBI, ExPASy, EBI, Israel, Japan]
Ilekis J.V., Stark B.C., Scoccia B.;
"Possible role of variant RNA transcripts in the regulation of epidermal growth factor receptor expression in human placenta.";
Mol. Reprod. Dev. 41:149-156(1995).

[3] SEQUENCE FROM NUCLEIC ACID (ISOFORM 2).
TISSUE=Placenta;
MEDLINE=97078686; PubMed=8918811; [NCBI, ExPASy, EBI, Israel, Japan]
Reiter J.L., Maihle N.J.;
"A 1.8 kb alternative transcript from the human epidermal growth factor receptor gene encodes a truncated form of the receptor.";
Nucleic Acids Res. 24:4050-4056(1996).

[4] SEQUENCE FROM NUCLEIC ACID (ISOFORM 2).
TISSUE=Placenta;
MEDLINE=97256547; PubMed=9103388; [NCBI, ExPASy, EBI, Israel, Japan]
Ilekis J.V., Gariti J., Niederberger C., Scoccia B.;
"Expression of a truncated epidermal growth factor receptor-like protein (TEGFR) in ovarian cancer.";
Gynecol. Oncol. 65:36-41(1997).

[5] SEQUENCE FROM NUCLEIC ACID (ISOFORMS 3 AND 4).
TISSUE=Placenta;
MEDLINE=21100872; PubMed=11161793; [NCBI, ExPASy, EBI, Israel, Japan]
Reiter J.L., Threadgill D.W., Eley G.D., Strunk K.E., Danielsen A.J., Schehl Sinclair C., Pearsall R.S., Green P.J., Yee D., Lampland A.L., Balasubramaniam S., Crossley T.D., Magnuson T.R., James C.D., Maihle N.J.;
"Comparative genomic sequence analysis and isolation of human and mouse alternative EGFR transcripts encoding truncated receptor isoforms.";
Genomics 71:1-20(2001).

[6] SEQUENCE OF 575-687 FROM NUCLEIC ACID.
Reiter J.L., Threadgill D.W., Danielsen A.J., Schehl C.M., Lampland A.L., Balasubramaniam S., Crossley T.O., Magnuson T.R., Maihle N.J.;
"Human and mouse alternative EGFR transcripts encoding only the extracellular domain of the receptor.";
Submitted (FEB-1999) to the EMBL/GenBank/DDBJ databases.

[7] SEQUENCE OF 713-924 FROM NUCLEIC ACID.
MEDLINE=84196372; PubMed=6326261; [NCBI, ExPASy, EBI, Israel, Japan]
Lin C.R., Chen W.S., Kruiger W., Stolarsky L.S., Weber W., Evans R.M., Verma I.M., Gill G.N., Rosenfeld M.G.;
"Expression cloning of human EGF receptor complementary DNA: gene amplification and three related messenger RNA products in A431 cells.";
Science 224:843-848(1984).

[8] SEQUENCE OF 150-962 FROM NUCLEIC ACID.
MEDLINE=84245835; PubMed=6330563; [NCBI, ExPASy, EBI, Israel, Japan]
Xu Y.H., Ishii S., Clark A.J.L., Sullivan M., Wilson R.K., Ma D.P., Roe B.A., Merlino G.T., Pastan I.;

http://ca.expasy.org/cgi-bin/niceprot.pl?P00533 Page 1 of 5

une autre BD

NicePro View of SWISS-PROT: P00533 1/10/03 9:43 AM

Cross-references

EMBL	X00588; CAA25240.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U95089; AAB53063.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48722; AAC50802.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48723; AAC50804.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48724; AAC50796.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48725; AAC50797.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48726; AAC50798.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48727; AAC50799.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48728; AAC50800.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U48729; AAC50801.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF288738; AAG35786.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF288738; AAG35787.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF288738; AAG35788.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF288738; AAG35789.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF288738; AAG35790.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF277897; AAK01080.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF125253; AAD30810.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	[illegible]	[EMBL / GenBank / DDBJ] [CoDingSequence]
	[illegible] JOINED.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	[illegible]	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X00663; CAA25282.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	M38425; AAA63171.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	M11234; AAA52370.1; -.	[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A00641; GQHUE. A00642; GQHUE2. A23062; A23062.	
HSSP	P11362; 1FGK. [HSSP ENTRY / PDB]	
SWISS-2DPAGE	P00533; HUMAN.	
Genew	HGNC:3236; EGFR.	
MIM	131550 [NCBI / EBI].	
GeneCards	EGFR.	
GeneLynx	EGFR; Homo sapiens.	
GlycoSuiteDB	P00533; -.	
SOURCE	EGFR; Homo sapiens.	
Ensembl	P00533; Homo sapiens. [Entry / Contig view]	
InterPro	IPR000494; EGFR_L_domain. IPR000719; Euk_pkinase. IPR002174; Furin-like. IPR001245; Tyr_pkinase. Graphical view of domain structure.	
Pfam	PF00069; pkinase; 1. PF00757; Furin-like; 1. PF01030; Recep_L_domain; 2.	
PRINTS	PR00109; TYRKINASE.	
ProDom	PD000001; Prot_kinase; 1. [Domain structure / List of seq. sharing at least 1 domain].	
SMART	SM00261; FU; 3. SM00219; TyrKc; 1.	
PROSITE	PS00107; PROTEIN_KINASE_ATP; 1. PS00109; PROTEIN_KINASE_TYR; 1. PS50011; PROTEIN_KINASE_DOM; 1.	
BLOCKS	P00533.	
ProtoNet	P00533.	
ProtoMap	P00533.	
PRESAGE	P00533.	
DIP	P00533.	
ModBase	P00533.	

Keywords

Transmembrane; **Glycoprotein**; **Repeat**; **Receptor**; **Signal**; **Transferase**; **Tyrosine-protein kinase**; **ATP-binding**; **Phosphorylation**; **Alternative splicing**.

Features

Key	From	To	Length	Description
SIGNAL	1	24	24	
CHAIN	25	1210	1186	EPIDERMAL GROWTH FACTOR RECEPTOR.
DOMAIN	25	645	621	EXTRACELLULAR (*POTENTIAL*).
TRANSMEM	646	668	23	*POTENTIAL*.
DOMAIN	669	1210	542	CYTOPLASMIC (*POTENTIAL*).
REPEAT	75	300	226	APPROXIMATE.
REPEAT	390	600	211	APPROXIMATE.

http://ca.expasy.org/cgi-bin/niceprot.pl?P00533 Page 3 of 5

EGF DANS PDB

Structure Explorer - 1FGK 1/10/03 9:22 AM

PDB PROTEIN DATA BANK

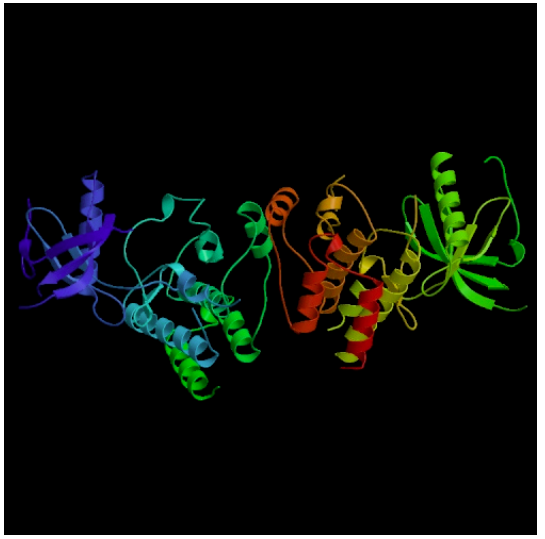
Structure Explorer - 1FGK

Title **Crystal Structure Of The Tyrosine Kinase Domain Of Fibroblast Growth Factor Receptor 1**
Classification **Phosphotransferase**
Compound **Mol_Id: 1; Molecule: Fgf Receptor 1; Chain: A, B; Fragment: Tyrosine Kinase Domain, Human Fgfr1 Residues That Possess Ptk Activity; Synonym: Fgfr1K, Fibroblast Growth Factor Receptor 1; Ec: 2.7.1.112; Engineered: Yes; Mutation: L457V, C488A, C584S**
Exp. Method **X-ray Diffraction**

View Structure

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Structure Factors (compressed)

Explore
SearchLite SearchFields



© *RCSB*

http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1FGK;page=&opt=show&size=500 Page 1 of 1

Séquences protéiques

Protéine : un ou plusieurs polymères d'acides aminés (polymère est une grosse molécule formée de l'union de molécules plus petites)

Alanine (Ala) Glycine (Gly) Méthionine (Met)

Sérine (Ser) Cystéine (Cys) Histidine (His)

Asparagine (Asn) Thréonine (Thr) Acide aspartique (Asp)

Isoleucine (Ilu) Proline (Pro) Valine (Val)

Acide glutamique (Glu) Lysine (Lys) Glutamine (Gln)

Tryptophane (Trp) Phénylalanine (Phe) Leucine (Leu)

Arginine (Arg) Tyrosine (Tyr)

La séquence détermine la structure

$\Rightarrow$ Problème de prédiction de structure protéique.

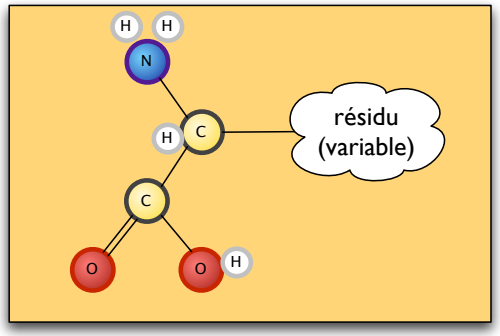
Acides aminés

[animation]

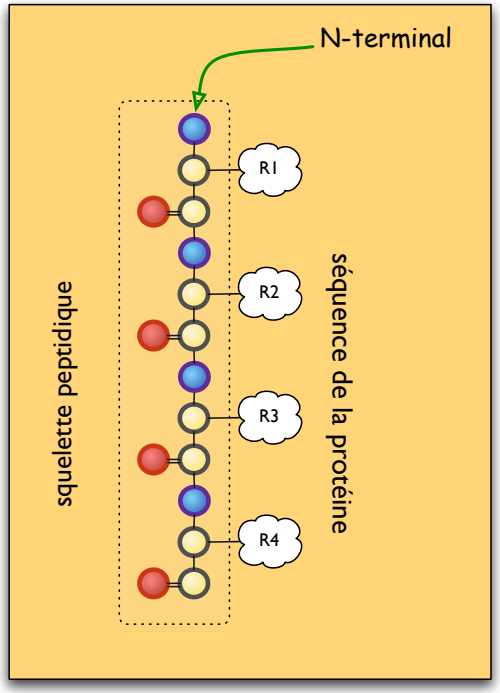
(Phe, Pro, Met, Gly)

Song Tan, Penn State U

Acides aminés et protéines



Acide aminé



Protéine

Autres problèmes

Comment est la structure établie ?

$\Rightarrow$ Repliement de protéines

La structure détermine la fonction

$\Rightarrow$ Problème de prédiction de fonction protéique.

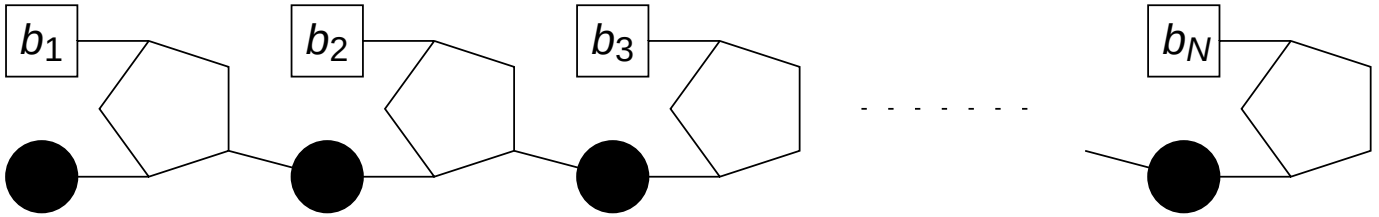
Interaction de protéines

$\Rightarrow$ Analyses de voies métaboliques, chemins de signalisation, réseaux régulatoires

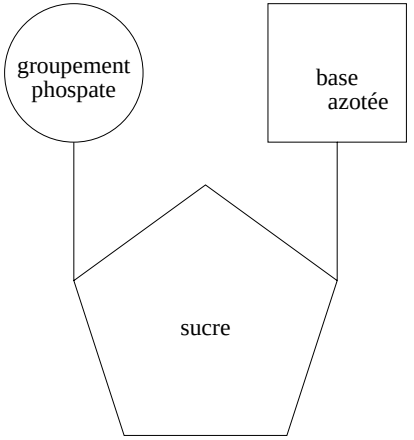
L'ADN ?

Polynucléotide :

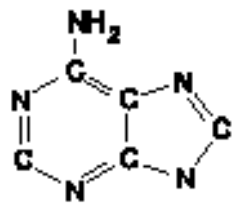
polymère (ie, chaîne) de nucléotides

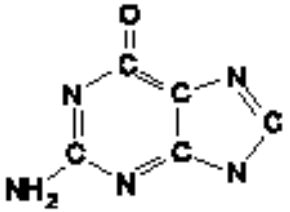


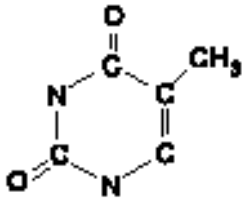
Nucléotides

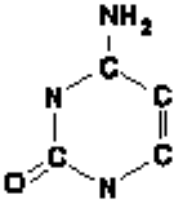


4 bases :

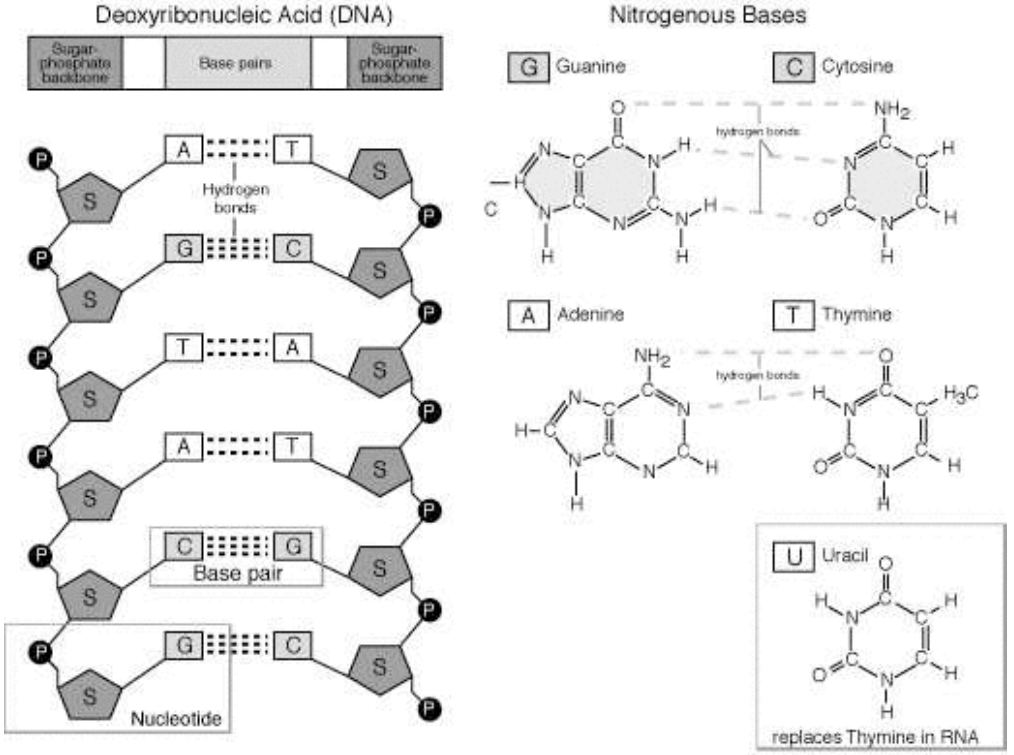

Adénine


Guanine


Thymine

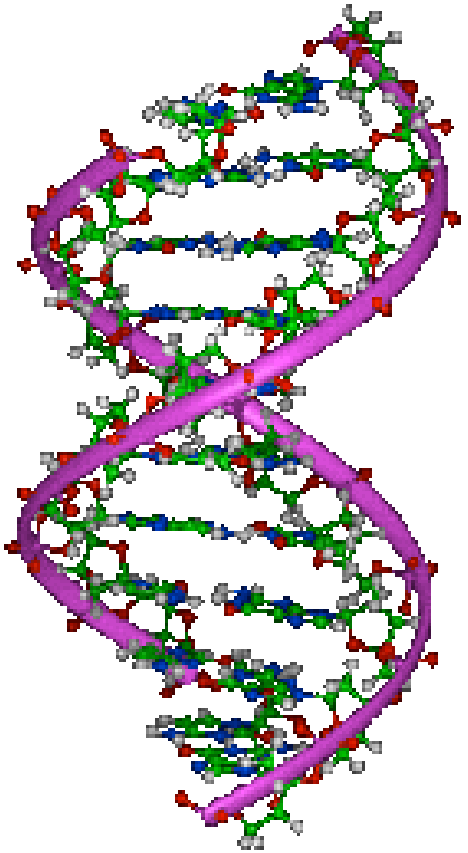

Cytosine

ADN : BASES COMPLÉMENTAIRES



Access Excellence

Watson et Crick : double hélice



1 pas de l'hélice : 10 paires de base, 3.4 nm (forme usuelle)

L'article

No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

1 Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).

2 Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).

3 Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (5) (1950).

4 Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (Stockholm), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the

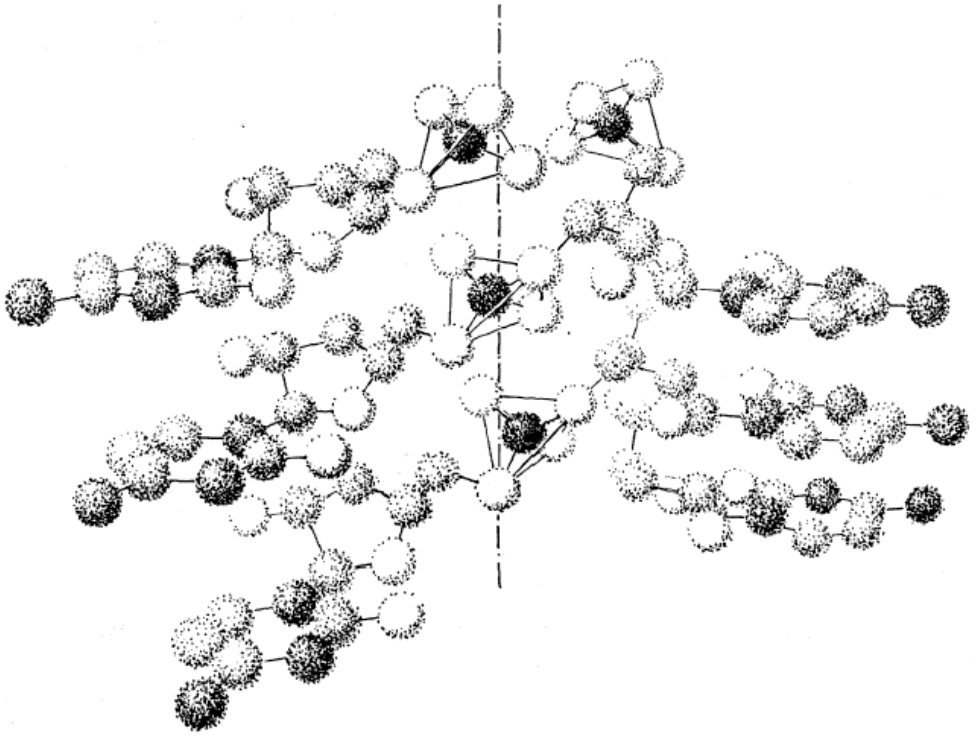
It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material

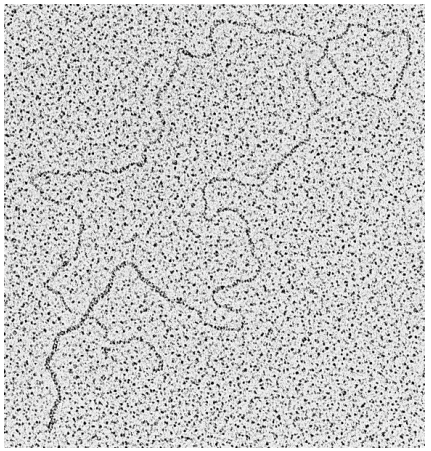
Un autre modèle — le triple hélice

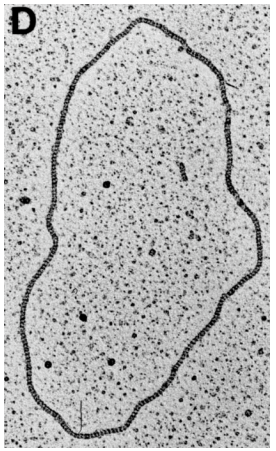


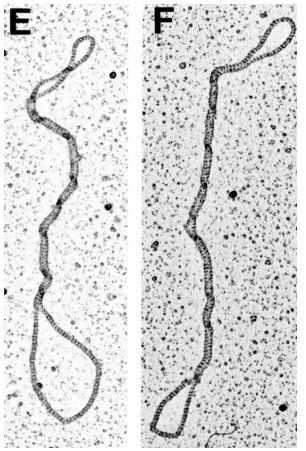
Pauling & Corey, *PNAS* 39 : 84 (1953)

ADN - STRUCTURE

la molécule d'ADN peut être
linéaire (nos chromosomes), ou
circulaire (bactéries, organelles)







Ross Inman

ADN

stockage de l'ADN : dans le noyau de la cellule, organisé en chromosomes (en eukaryotes)

1. duplication de l'information en ADN : hérédité
2. duplication de l'information en ARN (souvent traduit en protéines)

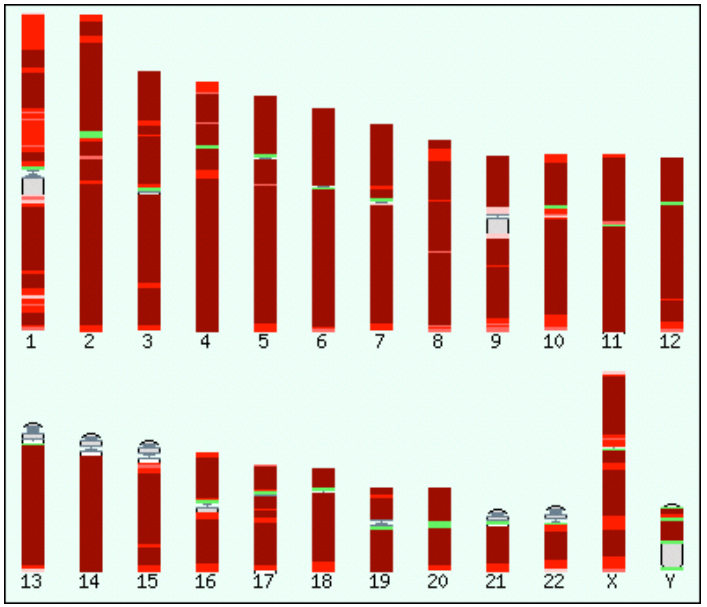
Replication

[animation]

[animation]

Lodish et al., *Molecular Biology of the Cell*, 2002

Chromosomes



22 chromosomes diploïdes et 2 chromosomes sexuels

tailles : $250 \cdot 10^6$–$23 \cdot 10^6$ paires de base
taille totale du génome humaine : $3 \cdot 10^9$ pb

NCBI

Structure de chromosomes et replication

[animation]

Dolan DNA Learning Center

Code génétique

20 acides aminés encodés par 4 nucléotides ?

Encodages par triplets : 64 codons

AAA = Phe	AAG = Phe	AAT = Leu	AAC = Leu
AGA = Ser	AGG = Ser	AGT = Ser	AGC = Ser
ATA = Tyr	ATG = Tyr	ATT = FIN	ATC = FIN
		...	

3 triplets d'arrêt, 1 triplet de début (encode aussi Met)

De l'ADN à protéine

1. transcription : copie du brin informatif à ARN messager
(ARN : utilise Uracile au lieu de Thymine)

2. traduction : ARNm à protéine (par le ribosome) : acides aminés fournis par ARN de transfert

Mécanisme universelle !
⇒ Problème de déterminer la séquence de l'ADN.

Gènes

Gène : unité d'hérédité, se traduit à une protéine (discussion plus raffinée plus tard)

Exons et introns
$\Rightarrow$ Problème de prédiction de gênes.

Expression d'un gène : quantité de ARNm transcrit $\Rightarrow$ Problème d'analyse de l'expression génique.

Comparaisons

Comparaisons de séquences : pour déterminer structure, fonction, ...

$\Rightarrow$ Problème de comparaison de séquences.

Grandes bases de données

$\Rightarrow$ Problème de recherche de séquences.

Autres sujets

Séquençage de l'ADN

Phylogénies

Calcul moléculaire