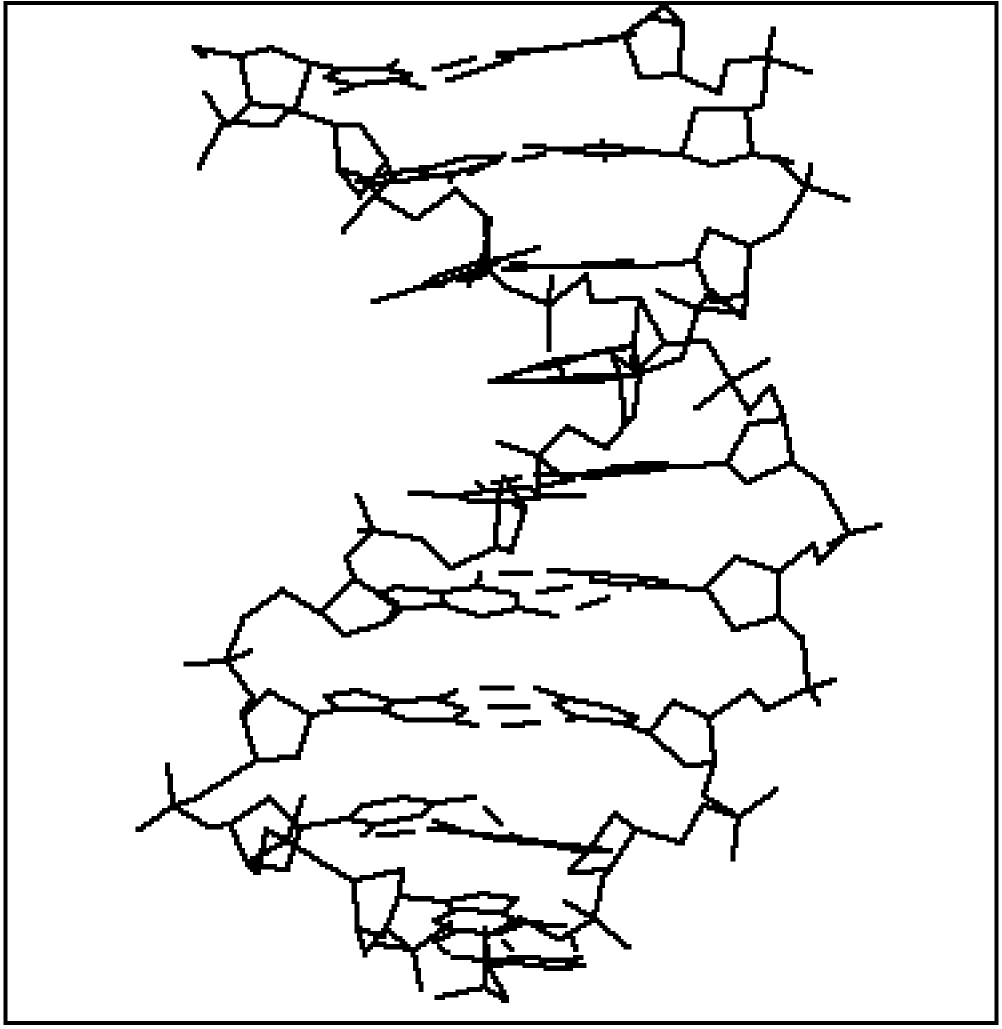
MOLECULAR BIOLOGY

DEOXYRIBONUCLEIC ACID (DNA)

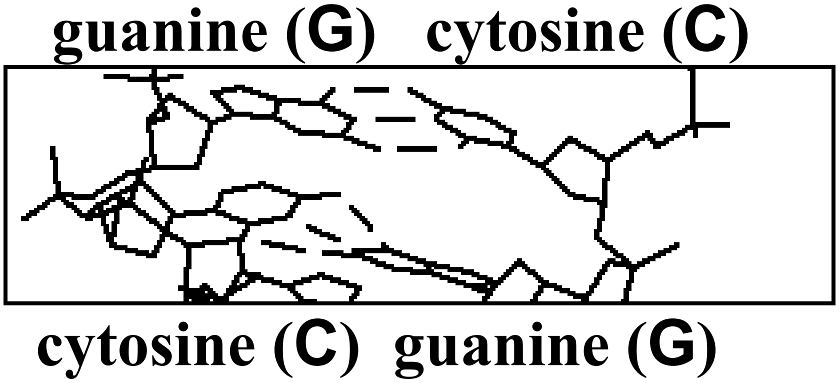
- **adenine (A)**
- **cytosine (C)**
- **guanine (G)**
- **thymine (T)**

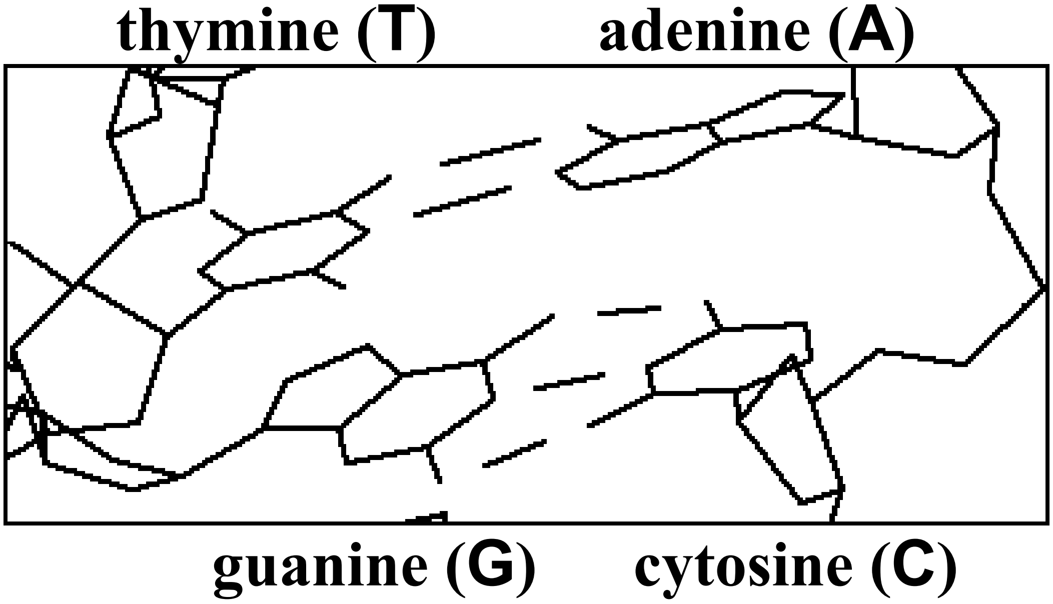


BASE PAIRING

DNA GGG TGC TCA

DNA CCC ACG AGT





TRANSCRIPTION (A --> U)

mRNA **GGG UGC UCA**

TRANSLATION

protein **G C S**

Gly Cys Ser

glycine cysteine serine

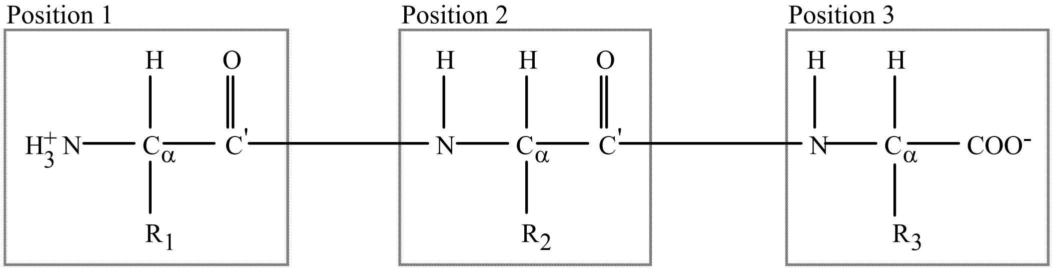
THE GENETIC CODE

One-letter code	Amino acid residue	Three-letter code	Genetic code
A	Alanine	Ala	GC*
C	Cysteine	Cys	UGU, UGC
D	Aspartic Acid	Asp	GAU, GAC
E	Glutamic Acid	Glu	GAA, GAG
F	Phenylalanine	Phe	UUU, UUC
G	Glycine	Gly	GG*
H	Histidine	His	CAU, CAC
I	Isoleucine	Ile	AUU, AUC, AUA
K	Lysine	Lys	AAA, AAG
L	Leucine	Leu	UUA, UUG,CU*
M	Methionine	Met	AUG
N	Asparagine	Asn	AAU, AAC
P	Proline	Pro	CC*
Q	Glutamine	Gln	CAA, CAG
R	Arginine	Arg	CG*, AGA, AGG
S	Serine	Ser	UC*, AGU, AGC
T	Threonine	Thr	AC*
V	Valine	Val	GU*
W	Tryptophan	Trp	UGG
Y	Tyrosine	Tyr	UAU, UAC

HEMOGLOBIN "S" (M25113 IN EMBL)

```
ATGGTGCACC TGACTCCTGT GGAGAAGTCY GCNGTTACTG CNYTNTGGGG    50
MetValHisL euThrProVa lGluLysSer AlaValThrA laXaaTrpGl
CAAGGTGAAC GTGGATGAAG TTGGTGGTGA GGCCCTGGGC AGGCTGCTGG   100
yLysValAsn ValAspGluV alGlyGlyGl uAlaLeuGly ArgLeuLeuV
TGGTCTACCC TTGGACCCAG AGGTTCTTTG AGTCCTTTGG GGATCTGTCC   150
alValTyrPr oTrpThrGln ArgPhePheG luSerPheGl yAspLeuSer
ACTCCTGATG CAGTTATGGG CAACCCTAAG GTGAAGGCTC ATGGCAAGAA   200
ThrProAspA laValMetGl yAsnProLys ValLysAlaH isGlyLysLy
AGTGCTCGGT GCCTTTAGTG ATGGCCTGGC TCACCTGGAC AACCTCAAGG   250
sValLeuGly AlaPheSerA spGlyLeuAl aHisLeuAsp AsnLeuLysG
GCACCTTTGC CACACTGAGT GAGCTGCACT GTGACAAGCT GCACGTGGAT   300
lyThrPheAl aThrLeuSer GluLeuHisC ysAspLysLe uHisValAsp
CCTGAGAACT TCAGGCTCCT GGGCAACGTG CTGGTCTGTG TGCTGGCCCA   350
ProGluAsnP heArgLeuLe uGlyAsnVal LeuValCysV alLeuAlaHi
TCACTTTGGC AAAGAATTCA CCCCACCAGT GCAGGCAGCC TATCAGAAAG   400
sHisPheGly LysGluPheT hrProProVa lGlnAlaAla TyrGlnLysV
TGGTGGCTGG TGTGGCTAAT GCCCTGGCCC ACAAGTATCA CTAAGCTCGC   450
alValAlaGl yValAlaAsn AlaLeuAlaH isLysTyrHi s...
TTTCTTGCTG TCCAATTTCT ATTAAAGGTT CCTTTGTTCC CTAAGTCCAA   500

CTACTAAACT GGGGGATATT ATGAAGGGCC TTGAGCATCT GGATTCTGCC   550

TAATAAAAAA CATTTATTTT CATTGC                             576
```

HYPOTHETICAL PROTEIN WITH UNSPECIFIED SIDE CHAINS



HYPOTHETICAL PROTEIN SEGMENT CONSISTING OF GLY, CYS, AND SER

THE STRUCTURE AND FUNCTIONS OF LIVING ORGANISMS ARE PRIMARILY DETERMINED BY PROTEINS

- **<u>Enzymatic catalysis</u>: Proteins catalyze chemical reactions in biological systems. Nearly all chemical reactions in biological systems are catalyzed by a specific macromolecule (i.e., an enzyme) and nearly all known enzymes are proteins.**

- **<u>Tranport and storage of ions and small molecules</u>: Examples: Myoglobin (stores oxygen), Hemoglobin (transports oxygen), transferrin (carries iron in blood), Ferritin (stores iron in liver).**

- **<u>Coordinated motion</u>: Examples: For muscle contraction, propulsion by flagella. Actin and myosin.**

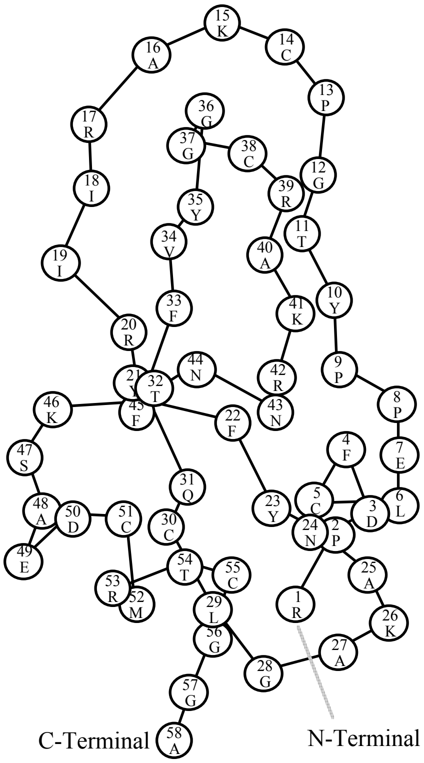
CONTINUATION - THE MANY ROLES OF PROTEINS

- **Mechanical Support: Example: the fibrous protein collagen.**
- **Immune protection: Example: Antibodies recognize and combine in highly specific ways with foreign entities such as bacteria. Self versus non-self.**
- **Generation and transmission of nerve impulses: Example: Rhodopsin is the photoreceptor protein in retinal rod cells and is used to generate nerve impulses.**
- **Control of growth and differentiation: For controlled sequential expression of genetic information. Examples: repressor proteins that silence portions of DNA, growth factor proteins, nerve growth factor proteins.**
- **Hormonal proteins: Transmit chemical instructions.**

PRIMARY STRUCTURE OF BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI)

```
RPDFCLEPPY TGPCKARIIR YFYNAKAGLC QTFVYGGCRA KRNNFKSAED   50

CMRTCGGA                                                 58
```

GENERAL STRUCTURE OF BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI)



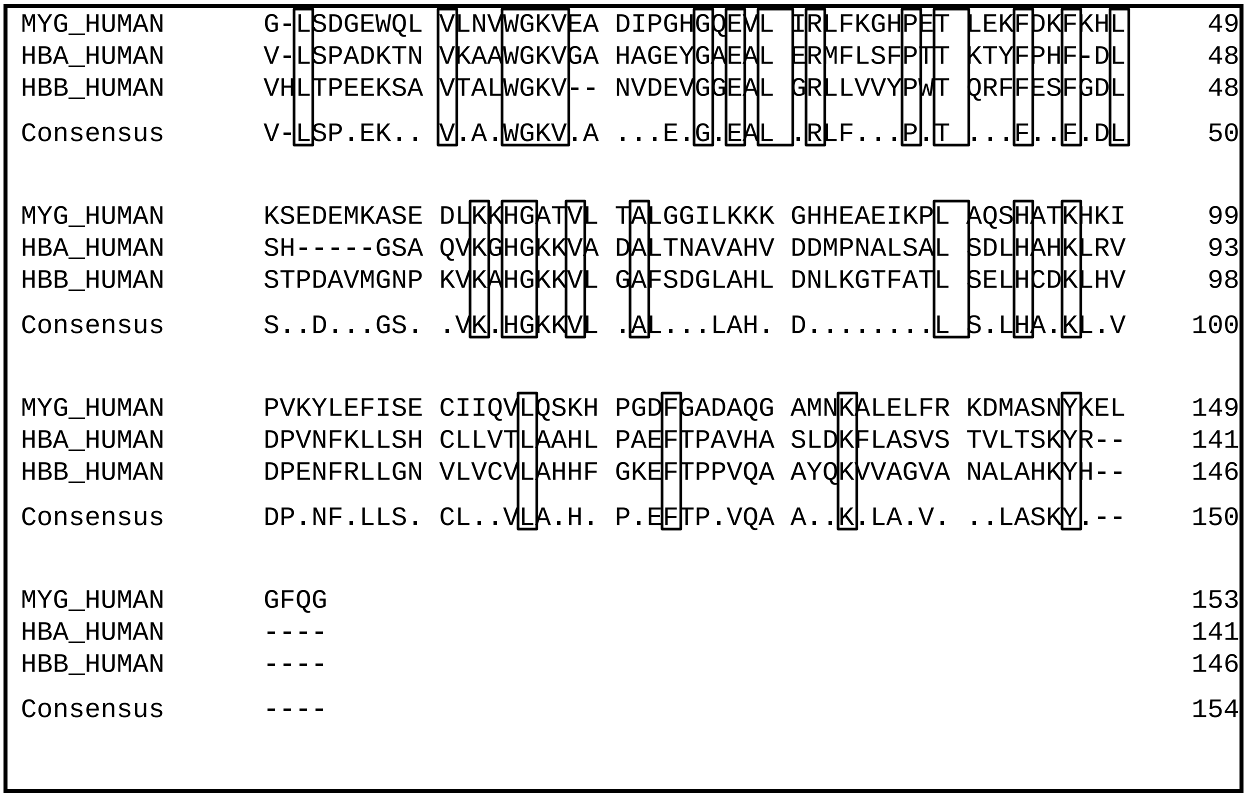
FEATURES OF THE SECONDARY STRUCTURE AND DISULFIDE BONDS OF BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI)

Feature	Type of feature	Start	End
H1	α-helix	Pro 2	Glu 7
H2	α-helix	Ser 47	Gly 56
S1	β-strand	Leu 29	Tyr 35
S2	β-strand	Ile 18	Asn 24
SS1	Disulfide bond	Cys 5	Cys 55
SS2	Disulfide bond	Cys 14	Cys 38
SS3	Disulfide bond	Cys 30	Cys 51

PORTION OF THE TERTIARY STRUCTURE OF BPTI FROM THE PROTEIN DATA BANK (PDB)

Amino acid residue	Residue number	Atom number	Atom	*x*	*y*	*z*
Cys	5	74	N	32.757	10.236	-6.732
Cys	5	75	α-C	31.286	10.029	-6.794
Cys	5	76	C	30.864	8.652	-7.254
Cys	5	77	O	29.690	8.279	-7.116
Cys	5	78	β-C	30.794	11.065	-7.789
Cys	5	79	γ-S	31.075	12.797	-7.325
Cys	5	80	D-H	33.206	10.888	-7.363
Cys	5	81	α-H	30.964	10.266	-5.800
Cys	5	82	β1-H	31.501	10.869	-8.603
Cys	5	83	β2-H	29.793	10.892	-8.171
Cys	55	883	N	28.364	15.919	-6.980
Cys	55	884	α-C	28.337	14.779	-7.839
Cys	55	885	C	27.258	14.663	-8.899
Cys	55	886	O	27.484	13.831	-9.733
Cys	55	887	β-C	28.265	13.520	-5.893
Cys	55	888	γ-S	29.664	13.161	-5.893
Cys	55	889	D-H	27.614	15.974	-6.323
Cys	55	890	α-H	29.253	14.775	-8.417
Cys	55	891	β1-H	27.388	13.519	-6.349
Cys	55	892	β2-H	28.059	12.720	-7.695

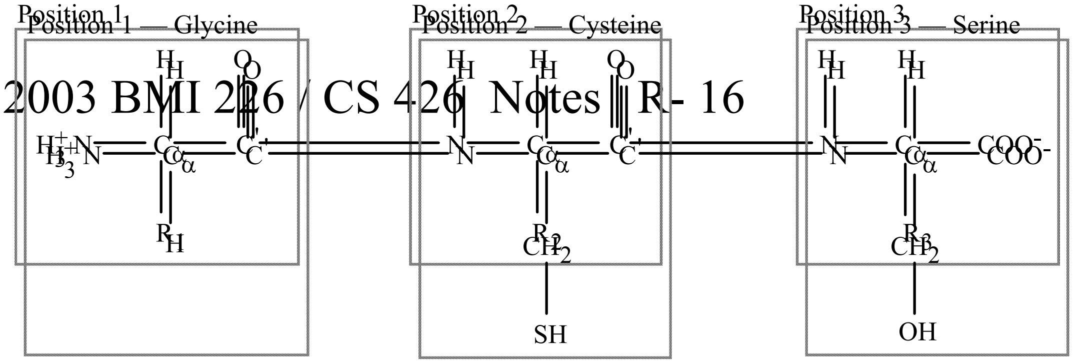
ALIGNMENT OF HUMAN MYOGLOBIN AND 2 CHAINS OF HEMOGLOBIN



```
MYG_HUMAN      G-LSDGEWQL VLNVWGKVEA DIPGHGQEVL IRLFKGHPET LEKFDKFKHL    49
HBA_HUMAN      V-LSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHF-DL    48
HBB_HUMAN      VHLTPEEKSA VTALWGKV-- NVDEVGGEAL GRLLVVYPWT QRFFESFGDL    48

Consensus      V-LSP.EK.. V.A.WGKV.A ...E.G.EAL .RLF...P.T ...F..F.DL    50


MYG_HUMAN      KSEDEMKASE DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI    99
HBA_HUMAN      SH-----GSA QVKGHGKKVA DALTNAVAHV DDMPNALSAL SDLHAHKLRV    93
HBB_HUMAN      STPDAVMGNP KVKAHGKKVL GAFSDGLAHL DNLKGTFATL SELHCDKLHV    98

Consensus      S..D...GS. .VK.HGKKVL .AL...LAH. D........L S.LHA.KL.V   100


MYG_HUMAN      PVKYLEFISE CIIQVLQSKH PGDFGADAQG AMNKALELFR KDMASNYKEL   149
HBA_HUMAN      DPVNFKLLSH CLLVTLAAHL PAEFTPAVHA SLDKFLASVS TVLTSKYR--   141
HBB_HUMAN      DPENFRLLGN VLVCVLAHHF GKEFTPPVQA AYQKVVAGVA NALAHKYH--   146

Consensus      DP.NF.LLS. CL..VLA.H. P.EFTP.VQA A..K.LA.V. ..LASKY.--   150


MYG_HUMAN      GFQG                                                     153
HBA_HUMAN      ----                                                     141
HBB_HUMAN      ----                                                     146

Consensus      ----                                                     154
```

GA'S AND PROTEIN FOLDING WITH SELF-AVOIDING GRAPHS

- **Unger, Ron and Moult, John. A genetic algorithm for 3D protein folding simulations. *Proceedings of the Fifth International Conference on Genetic Algorithms*. Ed. Stephanie Forrest. San Mateo, CA: Morgan Kaufmann Publishers, 1993. 581–588.**
- **Unger, Ron and Moult, John. “Genetic algorithms for protein folding simulations.” *Journal of Molecular Biology* 231 (1993): 75–81.**



UNGER AND MOULT – CONTINUED

HYPOTHETICAL PROTEIN WITH UNSPECIFIED SIDE CHAINS

HYPOTHETICAL PROTEIN CONSISTING OF GLY, CYS, AND SER

UNGER AND MOULT – CONTINUED

- **Individuals in the population are self-avoiding point-labeled (2 colors) graphs embedded in a 2-dimensional checkerboard lattice**
- **That is, individual in the population are the actual structures that the GA operates on**
  - **Phenotype (the individual) = Genotype**
- **2 psuedo-amino-acids:**
  - **Black (Hydrophobic)**
  - **White (Other)**

UNGER AND MOULT – CONTINUED

- **Fitness is decremented by -1 for each adjacent BLACK point along backbone that is not diagonally adjacent or adjacent along backbone**
    - **The 2 termini can contribute up to -3**
    - **Ordinary points can contribute up to -2**
- **There are 83,779,155 20-long self-avoiding graphs of the sequence. Fitness ranges from 0 to -9 (best) and there are only 4 9-scoring best conformations out of 83,779,155**

UNGER AND MOULT – CONTINUED

- **Mutation operation**
  - **Pick point**
  - **Keep trying random rotations that create self-avoiding graph as a result**
- **Crossover**
  - **Pick point**
  - **Keep trying random rotations that create self-avoiding graph as a result**

UNGER AND MOULT – CONTINUED

- **Population size *M* = 200**
- **Initialization: All alike (flat = 180 degrees)**
- **Accept result of mutation with Metropolis algorithm**
- **Accept result of crossover with Metropolis algorithm**
- **Global minimum of -9 found in all 5 runs after 8,800,000; 7,400,000; 3,200,000; 470,000; and 292,000 fitness evaluations. That is, between 9:1 and 284:1.**

SUN'S USE OF GA FOR PROTEIN TERTIARY STRUCTURE PREDICTION USING REDUCED REPRESENTATION MODEL

- **Sun, Shaojian. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Science*. Volume 2. Pages 762-785. 1993.**

- **Reduced representation**
  - **Only backbone atoms**
  - **Ideal fixed bond lengths and angles**
  - **Single virtual united-atom as side chain**

SUN – CONTINUED

- **Goal is to find the ϕ (phi) and ψ (psi) angles (2 per amino acid residue)**

- **Results in folded versions of**
  - **26-residue melittin – RMS error of 1.6 Å**
  - **36-residue avian pancreatic polypeptide inhibitor (APPI)**
  - **18-residue apamin (with 2 disulfide bonds) from bee venom**

SUN – CONTINUED

- **Fitness was a statistical interatomic potential function of his own design**
  - **Based on 110 proteins (with less than 50% identity)**
  - **melittin and avian pancreatic polypeptide inhibitor (APPI) were in the 110**
- **Fitness - 2 components**
  - **Local (NOTE: possible computer savings)**
  - **Non-local**
- **Apparently floating-point gene values. 2 x 26 = 52 for melittin. Values are integers from –180 to +180. Equivalent to 52 x 9 = 468 bits.**
- **Population size *M* = 90**

SUN – CONTINUED

Objective:	**Given the primary sequence of a protein, find the three-dimensional conformation of the protein in the form of the 2N dihedral ϕ and ψ angles using a reduced representation model of protein.**
Representation scheme:	**• structure = fixed length string (for a particular protein) • alphabet size *K* = 2 (in binary equivalent) • string length *L* = 468 (in binary equivalent) • mapping.**
Fitness cases:	**Only one (for a given protein).**
Raw fitness:	**Statistical fitness function.**

Parameters:	**• Population size$M$ = 90.** **• Maximum number of generations to be run$G$ = ???.** **• Special (???) mutation operation at ??? frequency**
Termination criteria:	**??? (Reports convergence of all 90!!!).**
Result designation:	**??? (Reports convergence of all 90!!!).**

SUN – CONTINUED

- **Reproduction NOT based on fitness. Creates 2M individuals.**

- **Crossover NOT based on fitness. Creates M individuals.**

- **Special mutation operation (sometimes changing several values at once). Creates 2M individuals.**

- **Selects the best M out of 5M new individuals.**

SUN – CONTINUED

- **On Gen 0, initial energy of 90 individual ranges from 1,440.08 to 15,746.34 units (with mean of 2912.00 and standard deviation of 1,960.75)**

- **On generation X, mean of the 90 individuals "converged" to 1,290.50 (with a standard deviation of 0.31 -– i.e., one part in about 4,000).**

LE GRAND'S USE OF GA FOR MINIZATION OF "AMBER" POTENTIAL ENERGY FUNCTION

- **Le Grand, Scott Michael.** ***The Application of the genetic algorithm to protein tertiary structure prediction*****. PhD Dissertation. Department of Biochemistry, The Pennsylvania State University, 1993.**

- **Goal is to find the two ϕ (phi) and ψ (psi) angles and 0-8 additional angles χ_1, ..., χ_8 per amino acid residue.**

LE GRAND – CONTINUED

- **Tried on 3 polypeptides**
  - **AGAGAGAGA (9 amino acid residues)**
  - **Polyalanine A9 (Alanine 9 times)**
  - **{Met}-enkephalin**

- **Tried on 4 proteins**
  - **46-residue crambin**
  - **26-residue melittin**
  - **36-residue avian pancreatic polypeptide inhibitor**
  - **106-residue cytochrome b562 (4 helix bundle)**

"AMBER" POTENTIAL ENERGY FUNCTION

Approximates N-body problem with 2-body terms by measuring all N^2 pairwise interactions of N atoms"

(1) VAN DER WAALS attraction and repulsion inversely depends on 12th and 6th powers of distance between each pair of non-bonded atoms. (Important at short range).

(2) COULOMB electrostatic attraction and repulsion inversely depends on 1st power of distance between each pair of non-bonded atoms. (Important at longer ranges).

"AMBER" POTENTIAL ENERGY FUNCTION – CONTINUED

(3) force (depending on square of deviation) to hold each 2-ATOM BOND DISTANCE at a constant equilibrium value. (Ignored by alternative functions that assume that bond length is fixed, except for disulfide bonds).
(4) force (depending on square of deviation) to hold each 3-ATOM BOND ANGLE at a constant equilibrium value. (Ignored by alternative functions that assume that 3-atom bond angle is fixed, except for disulfide bond angles).
(5) force is Fourier series with frequency and phase dependent on 4-ATOM DIHEDRAL ANGLE.

LE GRAND – CONTINUED

- **Fitness is AMBER plus additional van der Waals and Coulomb contributions for 1st and 4th atoms of 4-dihedrally-bound atoms AND additional van der Waals contribution for polar hydrogen and non-bonded oxygen and nitrogen.**
- **3 kinds of crossover (single-point, two-point, and uniform)**
- **steady-state GA. (Tends to be greedy).**
- **High (and changing) mutation rate.**
- **Child only replaces parent if it is better than most similar existing individual in the population (a variation of phenotypic sharing)**
- **Population size M = 200.**

LE GRAND – CONTINUED

Objective:	**Given the primary sequence of a protein, find the three-dimensional conformation of the protein in the form of the two ϕ (phi) and ψ (psi) angles and 0-8 additional angles χ_1, ..., χ_8 per each amino acid residue.**
Representation scheme:	**• structure = fixed length string (for a particular protein)** **• alphabet of real-valued genes**
Fitness cases:	**Only one (for a given protein).**
Raw fitness:	**AMBER-like potential energy function.**

Parameters:	**• Population size *M* = 200.** **• Maximum number of generations to be run specified as 100,000 (200 x 500) iterations.** **• Variation of phenotypic sharing.**
Termination criteria:	**100,000 (200 x 500) iterations OR variance of population is less than 0.1 OR average distance between 200 randomly selected pairs is less than 0.1.**
Result designation:	