Research Description

We use theoretical and computational techniques to help solve biological and medical problems. The current research topics can be grouped into the following five categories:

Protein structure and modeling We have a long-standing interest in the surface areas, volumes, cavities, stability and folding of protein structures. We served as one of three judges in the community-wide protein structure prediction experiment, CASP6, in 2004. Currently, we are exploring ways to utilize such structural information towards function determination. One way to gain functional information is to find a homologous protein for which such information is known. To this end, we have shown that the sensitivity of homology detection can be improved significantly by using minimal (residue burial) structural information.[REFERENCE LINK TO NALIN'S PAPER IN PRESS] In order to use structural information to improve homology detection in this manner, one needs a large number of accurate structure-based sequence alignments. We found that the accuracy of the structure-based sequence alignments produced by many structure comparison programs were disappointingly poor.[REFERENCE LINK TO CHANGHOON'S PAPER IN PRESS] We are currently working on ways to improve the accuracy of the structure-based sequence alignments. Since structurally similar proteins are likely to have a similar function, we are also studying the issues involved in the objective definition and automatic classification of protein structural domains.[REFERENCE LINKS TO TWO PAPERS WITH FRENCH COLLABORATION] We also work on obtaining structural models for a few specific protein molecules.

Immunotoxin Immunotoxins are man-made molecules constructed by joining an anticancer antibody and a suitable toxin, in our case, the *pseudomonas* exotoxin A. In all molecules under current active consideration, the antibody part is truncated to only the antigen-binding Fv portion of the molecule and the toxin part is modified to delete its own receptor-binding domain. Ideally, these molecules will bind only to the target cancer cells and kill them. Dr. Pastan's group in the Molecular Biology Section of our Laboratory has made many such molecules, each of which has a specific antibody for a particular cancer. Some of these have been tested in phase I clinical trials. In collaboration with this experimental group, we study the structural models of these molecules and attempt to find ways to improve their properties as an effective drug.[REFERENCE LINKS TO REITER ET AL. PROTEIN ENG. 1995; ONDA ET AL. J. IMMUNOL. 2006] We also construct mathematical models of immunotoxin delivery process in order to find ways to improve the efficacy of these agents.

Gene discovery We analyze the genome and expressed sequence (mRNA and EST) databases to discover genes that are specifically expressed in a particular organ or tumor. The products of such genes can potentially be used as targets for delivery of antitumor agents, for anticancer vaccine development, and for tumor imaging. In collaboration with Pastan's molecular biology group, we have found a number of such genes over the years, including *PAGE4, XAGE, PATE, MRP8, TARP, PRAC2, POTE, CAPC,* and *NGEP*. We also used these databases to discover novel fusion genes resulting from chromosomal

rearrangements, which are frequently involved in carcinogenesis.[REFERENCE LINK TO HAHN ET AL. PNAS 2004]

Comparative analysis of genes and genomes Comparison of human genes with their evolutionarily related homologs provides invaluable clues for the biological function of the proteins they encode. We collect, and attempt to construct the evolutionary history of, the homologs of the human genes identified by our Gene Discovery program. We found *ANKRD26* during this process, which is an ancestral gene of the *POTE* family of genes.[REFERENCE LINK TO HAHN ET AL. GENE, 2006] *ANKRD26* is in mouse, in contrast to the *POTE* genes which are primate-specific and absent in rodents. Pastan's group found that homozygous mice that carry an altered form of this gene become grossly obese. We also performed systematic searches for human-specific mutations that occurred after the Homo-Pan divergence by comparison of the human, chimpanzee and a third, outgroup, genome sequences.[REFERENCE LINKS TO HAHN'S 3 PAPERS: BIOINFORMATICS, 2005; HUM. GENET. 2006; MOL. BIOL. EVOL. 2007] The human-specific genetic alterations should be responsible for the generation of human-specific traits.

Hydrophobicity We study the phenomenon of hydrophobicity by means of statistical thermodynamics.[REFERENCE LINKS TO MY PAPERS, PNAS 1991; BIOPOLYMERS, 1991; PROT. SCI. 1993; MADAN & LEE, BIOPHYS. CHEM. 1994; LEE & GRAZIANO, JACS 1996; GRAZIANO & LEE, JPCB 2005] The hydrophobic effect is believed to be one of the main forces that determine the structure, stability, and interaction of protein and other biologically important molecules. This research is done in collaboration with Prof. Giuseppe Graziano at Università del Sannio, Benevento, Italy.

SUMMARY OF RESEARCH ACTIVITIES

Following are more specific descriptions of the research activity in the past four years.

PROJECT 1: PROTEIN STRUCTURE AND STRUCTURE MODELING

We have been studying the three-dimensional structure of protein molecules since the first few crystal structures were determined. We were the first to quantitatively define the solvent accessible surface of a protein molecule and to find cavities inside a protein molecule (1). We continue to study protein structures and to explore the space of protein folds. We devise and improve the methods for comparing and classifying them and for relating structures to their amino acid sequences. We also attempt to obtain structural models for a few specific proteins of interest.

Considering the sequence-structure relation first, finding the structure given a sequence is the protein structure prediction problem. Structure prediction presents a unique challenge to a computational physical chemist. In the past, our interest had been in *ab initio* prediction, in which one investigates how proteins fold on the basis of the first principles of physics and physical chemistry. However, it is the collective experience of the whole protein structure prediction community that *ab initio* techniques have yielded little

success so far whereas the 'knowledge-based' techniques, which rely on finding similar structures in the structural database, have been quite successful. We evaluated different structure prediction methods by participating in the CASP (Critical Assessment of Structure Prediction,) experiment in 2004. CASP is a well-known, public series of biennial experiments designed to objectively evaluate the state of the art of the structure prediction science (http://predictioncenter.gc.ucdavis.edu/). In these experiments, predictors worldwide submit models of proteins before the structures are known, which are evaluated by independent assessors after their experimental structures become available. We participated in the 2004 CASP6 experiment as one of the three assessors. This work confirmed that the methods that worked best were knowledge-based and that the first principle prediction of protein structure is as yet an unsolved problem. In addition, we could identify the methods that performed best and which could profitably be used for the structure prediction/modeling of specific proteins of our interest.

Another way to relate structure and sequence is to find sequences that will fit a structure. This is the reverse protein folding problem, but can be considered more generally as finding remotely homologous sequences using the structural knowledge of one member of the homologous family. Finding homologues is becoming increasingly more important because the most effective procedure for predicting the function of the product of a newly discovered gene is to find a homologous protein for which some functional information is available. Homology searches are also an essential step in establishing phylogenetic relations among different genes. Homology searches usually require using a tool such as BLAST (2) to identify sequence alignments with sufficiently high score. Obviously, the sensitivity and specificity of the search depends critically on the score matrix used. The score matrices commonly used today (3, 4) are based on amino acid substitution frequencies derived from sequence alignments alone; the structural information is not used even when such information is available for some of the homologous proteins. However, it is well known that amino acid substitution patterns depend heavily on the structural context. For example, an amino acid is most likely to be substituted by a non-polar residue if it is buried in the protein structure, but by a polar residue if exposed to the solvent. When many homologous sequences can be found using a conventional score matrix, then a position-specific score matrix (PSSM or profile) can be set up, which implicitly includes structural information. The PSI-BLAST program (5), which builds and uses PSSM in iterative fashion, greatly extends the power of BLAST to find more homologous sequences. However, when many homologous sequences cannot be found by using the conventional score matrix, or when all sequences found are highly sequence-similar, an effective profile cannot be constructed and PSI-BLAST loses its power. One may expect that sensitivity and specificity of the search would increase in such cases if the structural context effect were included directly in the amino acid substitution score matrix. We proved that such is indeed the case by developing the Context-Specific Score Matrices (CSSM) and demonstrating their power in finding more homologous sequences once the structure of one protein is known.

In order to obtain an overview of all protein structures that exist, the structures must be compared to each other and classified. Many, including us, have recognized the importance of objective structure comparison/alignment and written computer programs

for it. These programs are essential for detecting commonalities and differences among protein structures and for protein structure classification. They also produce structure-based sequence alignments, which are used as the gold standard of sequence alignments. Unfortunately, there are many such programs and they do not produce identical results. We have investigated the accuracy of the structure-based sequence alignments that some of these programs produce against an expert-curated alignment database. We have also used some of these programs to investigate the nature of the space of protein folds and the effect of structure clustering algorithms on protein structure classification.

We attempt to obtain structural models of the product of the genes that we newly discover and of other proteins of interest to other members of the laboratory. Structural models often provide information on possible function of the protein. Modeling exercises are useful to us also because they provide an opportunity to evaluate in real life situations the tools that we and others develop, e.g., tools for structure prediction, homology search, sequence alignment, and structure comparison, among others.

Specific Research Aims

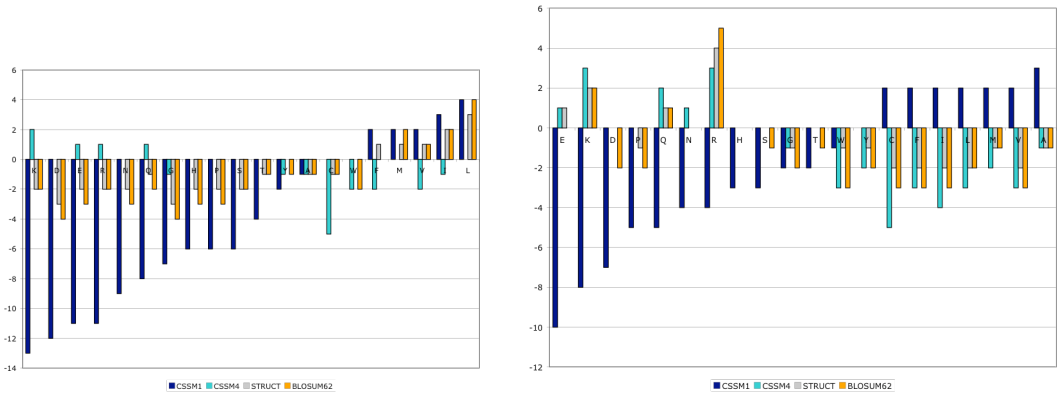
Sub-project 1. To enhance the sensitivity and specificity of homology searches through the protein sequence database by using structural information.
Sub-project 2. To evaluate structure and domain prediction methods by assessing their performance in the CASP6 experiment.
Sub-project 3. To evaluate the accuracy of structure-based sequence alignments that different structure alignment programs produce and to design new algorithms that will improve the performance of the structure comparison programs.
Sub-project 4. To explore the properties of the protein fold space by comparing automatically generated set of protein similarity/dissimilarity measures to expert-curated protein classification database.
Sub-project 5. To obtain three-dimensional structural models of specific proteins.

Accomplishments

Sub-project 1. In order to obtain amino acid substitution matrices that use structural context, we first set up a structure-based sequence alignment database, called SHoPP, by running our own structure alignment program SHEBA (6) on all pairs of domains in the ASTRAL SCOP v1.59 (40% ID) protein domain database (7) and selecting pairs that were structurally aligned. Then we built the CSSM matrices from the amino acid substitution frequencies that are observed in the SHoPP database, one matrix for each of the 4 degrees of burial of the residue substituted. A sample of some of the matrix elements is shown in Figures 1A and 1B. It can be seen that these matrices are strikingly different from the commonly used BLOSUM62 matrix (3, 4). These matrices were implemented in BLAST and PSI-BLAST programs so that they can be used for routine homology searches and for building the PSSM (5), when the degree of burial information is available for the sequence of interest. The ROC curves of true versus false positive hits (Figure 1C) show that the BLAST and PSI-BLAST that use the CSSM clearly outperform those that use the traditional BLOSUM62 matrix. Dr. Nalin Goonesekere,

who worked on this project while at NIH, is now at the University of Northern Iowa, where he calculated the PSSM profiles for all the chains in the protein structure database (PDB) (8) using the CSSM-PSI-BLAST. His preliminary studies indicate that one can associate many sequences (those in nearly 300 Pfam (9) DUF (Domain of Unknown Functions) families) to a PDB structure for the first time by using RPS-BLAST to search through this profile database. We are in the process of setting up a web server to let others have free access to this operation. A manuscript describing this work has been accepted for publication in the Proteins and is attached to this report.

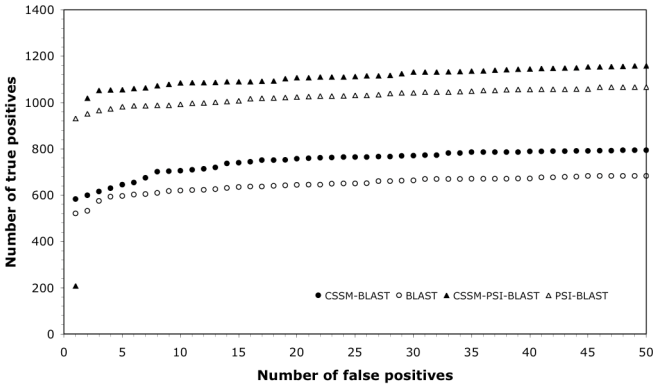
Figure 1.



(A) (B)

(A) Leucine to other amino acid substitution scores in $CSSM_1$ (black), $CSSM_4$ (green), STRUCT (grey), and BLOSUM62 (orange) score matrices. $CSSM_1$ and $CSSM_4$ matrices are for substitution of residues in the buried (< 25% solvent accessible) and exposed (> 75% solvent accessible) positions, respectively. STRUCT is the substitution score matrix built as the CSSM matrices from the same SHoPP database, but using pooled frequencies ignoring the solvent exposure. BLOSUM62 is a popular score matrix built from multiple sequence alignment data (3, 4). The substituting amino acids are sorted in ascending order of the $CSSM_1$ score. It can be seen (a) that the values for the $CSSM_1$ and $CSSM_4$ matrix elements often have opposite signs for substitution by the same amino acid types, (b) that the values for the STRUCT and BLOSUM62 tend to be in between the $CSSM_1$ and $CSSM_4$ values, and (c) that STRUCT and BLOSUM62 values are similar, indicating that the difference we see between CSSM and BLOSUM62 is not primarily due to the database difference. **(B)** Similar to (A), but for the substitution of the arginine residue. Notice that the score pattern is similar between panels (A) and (B) for the CSSM matrices, but quite different for the STRUCT and BLOSUM62.

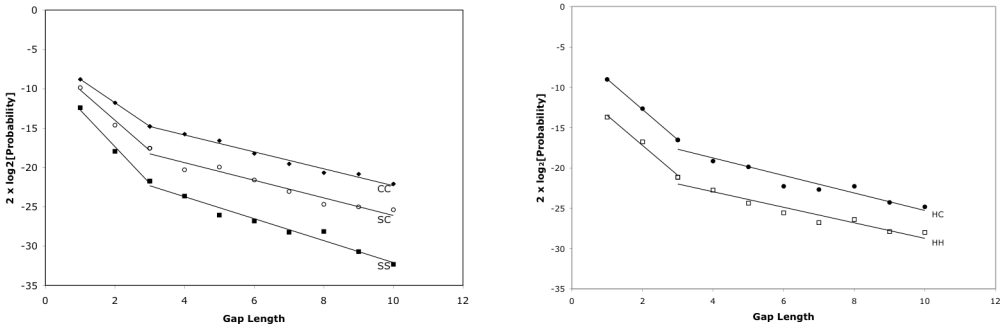
(C) ROC_{50} curves computed from pooled results for a set of 92 query sequences with ASTRAL-SCOP v1.65 (50% seq. ID) as target database, using CSSM-BLAST (solid circles), regular BLAST with BLOSUM62 (open circles), CSSM-PSI-BLAST (solid triangles) and a default version of PSI-BLAST using BLOSUM62 (open triangles). Hits to the same SCOP superfamily were considered true positives. Hits to different SCOP folds were considered false positives.



(C)

An effective homology search by sequence alignment also requires a good gap penalty function. We investigated the effect of structure in the frequency of the gaps that occur in the SHoPP database and found (a) that the frequency depended strongly on the secondary structural type of the residues flanking the gap, which was not surprising, and (b) that, when the frequencies were examined separately for each secondary structural type, the logarithm of the frequencies varied linearly with the gap length, but with a pronounced break in the slope at gap length of 3, which was unexpected (Figures 2A and 2B). We proposed a new gap penalty function on the basis of these observations. Use of such context-dependent gap penalties should produce a better alignment. This work has been published (10).

Figure 2.



(A) (B)

Probability distribution of gaps, by length, for data from the SHoPP Database. Gaps have been categorized by the secondary structure of the residues flanking the gap.
(A) CC - Gaps within a coil; SC - Gaps at the edge of a strand; SS - Gaps within a strand.
(B) HH - Gaps within a helix; HC - Gaps at the edge of a helix.

Sub-project 2. In the 2004 CASP6 experiment, there were 63 target proteins of unknown structure. The three teams of assessors (B. Lee at NIH, R. Dunbrack at Fox Chase Cancer

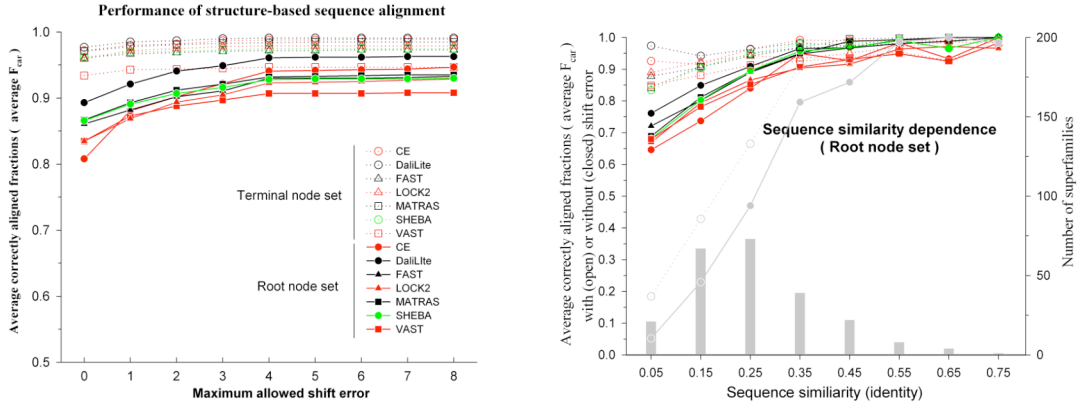
Center, Philadelphia, PA and A. Valencia at Centro Nacionale de Biotecnologia, Cantoblanco, Spain) visually examined each of these targets, after their structures became known, and parsed them into 90 domains. The domains were then classified (11) into New Fold (NF) targets, for which no similar structure existed in PDB, Fold Recognition (FR) targets, for which similar structures existed in PDB but they had either no detectable (FR/A) or barely detectable (FR/H) sequence similarity, and Comparative Modeling (CM) targets, for which existing sequence homology search programs can find a similar structure in PDB. We took the responsibility of evaluating models for the 9 NF targets and 8 of the 16 FR/A targets with low structural similarity to a known structure. There were over 7400 models for these 17 targets, which were submitted by 165 teams of researchers all over the world. We started evaluating them as soon as the true structures became available. Initially, the organizers of the CASP6 experiment automatically assigned a numerical score to each model by using the program LGA (12), which had been devised especially for detecting local and global similarities between a model and the experimental structures. We then visually inspected high scoring models for each target, the number of which varied from 20 to over 100 depending on the target. We also calculated average scores for each prediction group for the purpose of ranking them. We found that models that bore overall similarity to the true structure were submitted for 7 of the 8 FR/A targets, but only for 3 or 4 of the 9 NF targets. High scoring models were submitted by several different groups; the single group that was most successful was Baker's group at the University of Washington, who submitted best models for 7 of the 17 targets. We also found what others had found before us, that the most successful prediction methods work by copying existing structures well and that truly *ab initio* methods were not among the most successful (13).

We also evaluated domain boundary predictions, which were introduced for the first time in the 2004 CASP6 experiment. Correctly predicting domain structure of a protein is obviously important for a successful structure prediction and also for protein engineering tasks wherein one wants to remove or replace a domain. For this purpose, we devised a new scoring scheme, called the Net Domain Overlap (NDO) scores, which is based on giving equal weight to the reward for correctly predicted residues (true positives) and penalty for the incorrectly predicted residues (false positives). In the past, domain predictions were measured by means of two or more different measures, typically including the difference in the number of predicted and actual domains and the number of residues between the predicted and actual domain boundaries. But true number of domains is sometimes not known because some part of the protein is missing in the crystal structure. It is also difficult to come up with a ranking of the predictions if more than one measure are used because they usually have different units and there is no guidance on how to scale or give a relative weight to one against the other. The new scoring function shows good correlations with both of these measures and can effectively replace them (14). The NDO scoring scheme was adopted by the CASP7 assessors as a part of the 2006 CASP7 domain prediction evaluation (http://predictioncenter.org/casp7/meeting/presentations/Presentations_assessors/CASP7_DP_Tress.pdf). The NDO scores, or a suitably modified form of it, should be generally useful in comparing any two partitions (sets of clusters) of objects, like the domains (clusters of residues) in a protein structure, groups of protein structures in the protein fold

universe, or phylogenetic clusters of genes. We used a modified version of NDO scores to assess the quality of antibody clusters based on competitive binding assays (see project 3). Returning back to CASP6, we found that the predictors used many different strategies to predict the domain structure of a protein. Most one- or two-domain proteins were predicted quite accurately, but the predictors had difficulty when the number of domains was more than two or when domains were made of more than one non-contiguous segments (14).

Sub-project 3. We used 7 easily available pair-wise structure alignment programs [CE (15), DaliLite (16), FAST (17), LOCK2 (18), MATRAS (19), SHEBA (6) and VAST (20, 21)] to structurally align a set of protein pairs prepared from NCBI's Conserved Domain Database (CDD) (22). The alignments generated by these programs were then compared with the CDD alignments, which were expert-curated. The degree of agreement was measured by the fraction of correctly aligned residues, $f_{car}(\delta)$, where the alignment of a residue was considered 'correct' if it aligned to a residue that is within δ residues from the residue it is aligned to in the reference alignment. The $f_{car}(0)$ values are useful for measuring absolute alignment accuracy, which is needed for example for obtaining accurate amino acid substitution profiles, while $f_{car}(8)$ values measure the fraction of the structure recognized as being similar without insisting on high resolution accuracy, which are useful for applications such as structure recognition and classification. We found that 4 to 9% of the CD core residues on average were either not aligned or aligned with more than 8 residues of shift error and that an additional 6 to 14% of the conserved residues on average were misaligned by 1-8 residues, depending on the program and the dataset used (Figure 3A). The alignment accuracy depended on sequence similarity even though the programs aligned solely or primarily on the basis of the geometric structure (Figure 3B). Also, there was a large variation in the alignment accuracy depending on the protein pair and on the structure comparison programs (Figure 3C).

Figure 3.



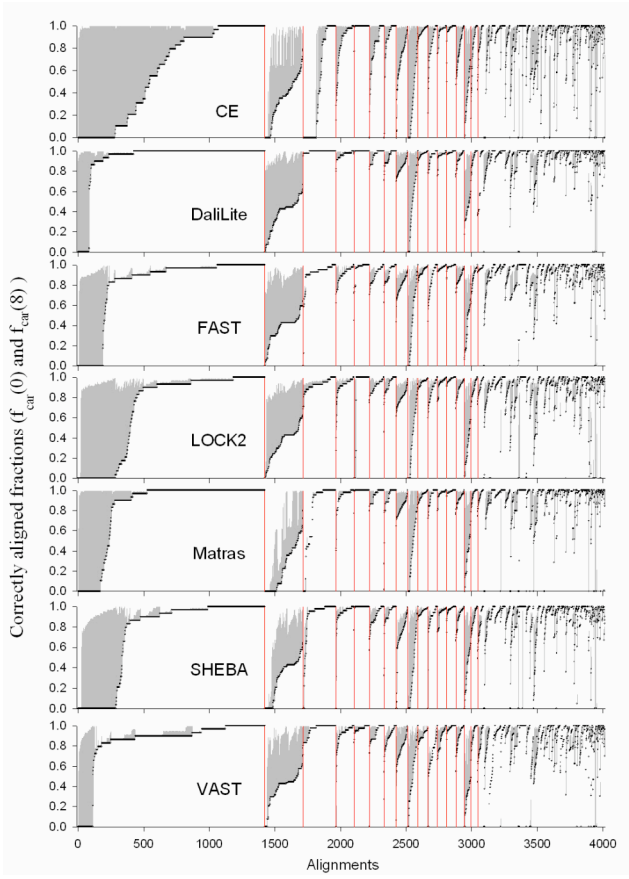
(A) (B)

(A) Average F_{car} (CD family-wide average of f_{car}) as a function of the magnitude of the allowed shift error. CDD is a hierarchically arranged database of multiple alignments of conserved groups of protein domains. We culled two datasets from this database, the root and the terminal node sets. The root node set consists of alignment pairs in the top-most

(most remotely homologous) groups in the CDD hierarchy whereas the terminal node set consists of pairs that are in the lowest group (most similar). The terminal and the root node sets are indicated by dotted lines with open symbols and solid lines with closed symbols, respectively. Program names are given in the alphabetical order. Note that the y-axis scale is from 0.5 to 1.0.

(B) Sequence similarity (fraction of identical pairs) dependence of F_{car} in the root node set. Alignments were grouped into sequence similarity bins of size 0.1 and then the alignments within each bin were grouped according to its CD name for averaging. The average F_{car} values are shown with the scale on the left y-axis: open symbols, $F_{car}(8)$; closed symbols, $F_{car}(0)$. The x-axis shows the midpoint of each sequence similarity bin. Different methods are indicated by different colors using the same color scheme as in (A). The gray circles and lines show the alignment accuracy obtained by SSEARCH (23), which is a purely sequence alignment program, provided here for contrast. The figure shows that structure-based alignments are clearly better than pure sequence alignment when the sequence identity falls below about 50%. The histogram (grey bars) shows the number of superfamilies in each bin with the scale on the right y-axis.

(C) The fraction of correctly aligned residues (f_{car}) of each alignment for each method**.** The superfamilies along the x-axis are sorted in descending order of the size of the superfamily (number of alignment pairs in the superfamily). Boundaries of the large superfamilies (those with 50 or more alignment pairs) are marked by red vertical lines. The alignments in each superfamily are sorted in ascending order of $f_{car}(0)$, which are shown as black circles. The grey vertical lines cover the range between $f_{car}(8)$ and $f_{car}(0)$ for each alignment. The methods are given in alphabetical order. Note that the order of superfamilies along the x-axis is preserved for all methods, but the order of the individual alignments within a superfamily is not since they are sorted by $f_{car}(0)$ values, which are specific for each method. Superfamilies marked by the red boundary bars are, from left: cd00096, **cd02156**, cd01983, cd00900, cd00657, cd02019, cd03440, cd01292, **cd02688**, cd00314, cd00196, cd00650b, cd00650a, cd00768, **cd02184**, and cd00267. The bold-faced superfamilies are those for which the $f_{car}(0)$ values are low (longest grey lines) for all methods.



A manuscript describing this study is under review for publication in BMC Bioinformatics and is attached to this report. Our own structure alignment program,

SHEBA, is not a particularly high performer according to this evaluation and must have adversely affected the quality of the SHoPP database that we used to set up the CSSM-based profiles (see above). We will replace the SHoPP database and the profiles when we develop an improved structure alignment program. (See Current and Future Research.)

Sub-project 4. We used two structure comparison programs, SHEBA (6) and VAST (20, 21), to calculate all-against-all pair-wise similarity scores and compared them to the similarities implied by the manually curated SCOP Fold classification (24). We found that the agreement was poor; at 1% false positive rate, only 62%-75% of the pairs within the same SCOP folds were considered similar by the automatic programs. We could identify four major causes of such discrepancy between machine-calculated and human-curated similarities: (a) the sub-structural feature that is considered to be the common core of a set of proteins by human experts can vary significantly and be judged to be different by automatic computer programs, (b) the common core of a SCOP Fold is too small a fraction of the total structure, which contains many other residues that are not structurally similar, (c) structures in different SCOP Folds may contain similar sub-structures, which produce false positives, or (d) some SCOP Folds consist of variable number of repeating units and include proteins with greatly differing sizes. This work was a collaboration between our group, Dr. Munson's group at CIT, and Drs. Garnier and Gibrat of the INRA, Jouy-en-Josas, France, and has been published (25).

Modifying the structure comparison/classification methods will probably reduce some or most of these differences. But it is possible that some of these arise from the continuous and multi-dimensional, rather than discrete and one-dimensional, nature of the protein fold space. When protein structures vary continuously in more than one aspect, and similarity is decided by using a cutoff value, the similarity property is not necessarily transitive, i.e., if structures A and B are judged to be similar and B and C also, it does not necessarily follow that A and C will also be judged to be similar using the same cutoff value. On the other hand if the structures are classified into discrete folds, then transitivity is a necessary property, i.e. if A and B are in the same cluster and B and C also, then A and C are necessarily in the same cluster. In order to see how much of the discrepancy that we observed in above study is due to such effect of classification, we classified protein structures using the machine-calculated all-against-all pair-wise similarity scores. We investigated a great many different clustering procedures. We also used the concept of homogeneous clusters to bring out irreducible differences between machine and human-curated SCOP partitions. These efforts produced convincing evidence that the major portion of the discrepancy is due to the difference in the way that similarity is perceived by machine and human and not to the effect of clustering. This study was also a result of the collaboration of the same teams. A manuscript has been submitted for publication and is attached to this report.

Sub-project 5. Using structure prediction and other bioinformatics tools, we were successful in obtaining some structural information for the protein products of all the genes that we discovered in the gene discovery project, including two (NGEP and CAPC) that we published in the past four years. (See the report on the Gene Discovery project.)

RepA is the P1 plasmid-encoded protein that binds to the plasmid's replication origin to initiate DNA replication (26). Its biological function is similar to that of RepE of the F plasmid, for which a crystal structure has been determined (27). The two proteins are related since both belong to the Pfam (28) Rep3 family but the sequence similarity is low and the BLASTP program does not find one from the other. We aligned RepA to RepE using two fold recognition programs, bioinbgu (29) and 3D-PSSM (30), and combined the alignment with the previously published multiple alignment of RepE and the initiator proteins of R6K, pSC101, pCU1, and pPS10 (27). The alignment was then manually adjusted to eliminate breaks or insertions in secondary structures compared with the x-ray structure of RepE. The final alignment was used to build a model of the monomer of RepA using the crystal structure of RepE as the template and the modeler program in the Insight II program package (Accelrys Software Inc, San Diego, CA, USA). Dr. Wickner's group in this laboratory tested the model by introducing mutations in the two predicted DNA binding regions and verified that the mutants were indeed defective in P1 DNA binding in the predicted manner (31).

Current Research and Future Plans

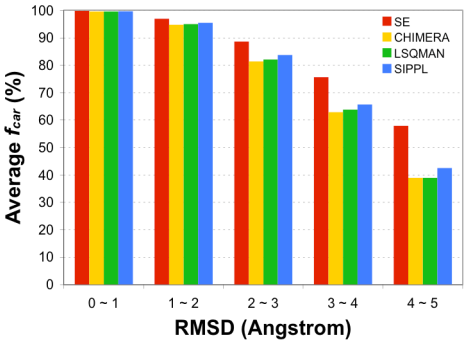
(1) Set-up and maintenance of the RPS-BLAST web server for profile search: We are currently in the process of setting up a web server to run the RPS-BLAST through the CSSM-based profiles that we calculated for each chain in PDB. We would like to finish this process and maintain the server so that we and others can make routine RPS_BLAST runs through the new profiles. We would also like to update the profiles periodically as the PDB gets updated and as new structure alignment algorithms are developed. This will be done in collaboration with Dr. Nalin Goonesekere of the University of Northern Iowa.

(2) New structure comparison/alignment algorithm: We have seen the need for accurate structure-based sequence alignment in two occasions as described above, once for setting up accurate CSSM and second time to obtain accurate measures of structural similarity for structure classification. In the former case, $f_{car}(0)$ should be high since accurate pair-wise alignment in the SHoPP database is essential. In the latter case, perhaps a fine alignment accuracy is not necessary, but the structure comparison program must still produce highly accurate measure of the degree of similarity, i.e. $f_{car}(8)$ must be high for similar structures. Yet, we have seen that current structure alignment programs produce errors for about 5% of the conserved core residues on average as measured by $f_{car}(8)$ and about 15% on average as measured by $f_{car}(0)$. There are many individual cases wherein the f_{car} value dips below 80%. By carefully analyzing the results of 7 different programs, we have also learned the strengths and weaknesses of each program and, in some cases, why they fail.

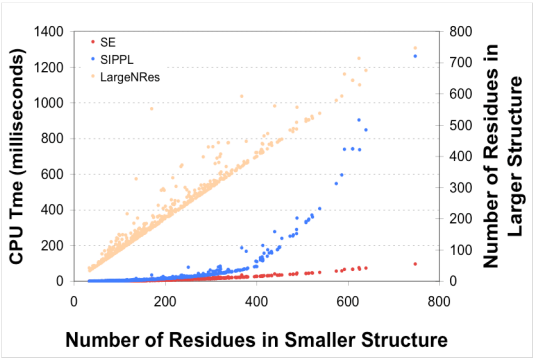
SHEBA, like many other similar programs, works in two stages; first it finds an initial alignment, which is iteratively refined in the second stage. Errors and imperfections can occur in both of these stages. We addressed the problems that occur in the second stage first. This stage is a two-step cycle of optimal structure superposition for the given alignment followed by finding new alignment from the superposed structures. There is an elegant mathematical solution for obtaining optimal superposition when an alignment is

given (32, 33) and there is no room for improvement in this step. On the other hand, finding the optimum sequence alignment from two superimposed structures is a non-trivial problem and some of the inaccuracies of the structure-based sequence alignment arise from poor performance of this part of the structure alignment programs. Many programs, including SHEBA, use dynamic programming algorithm with a gap penalty for this purpose. Unfortunately, there is no guidance on proper gap penalties to use since gaps of all sizes, including a whole domain, occur in protein structures. SHEBA uses the gap penalty value of 0, which results in too many gaps and incorrect alignments especially in comparing two helical domains. We devised a new algorithm for this step, which works somewhat like the BLAST algorithm in that it finds short seed alignments, which are then extended and merged to obtain one consistent overall alignment. The new algorithm, which we call SE for Seed Extension, does not use gap penalty and significantly improves the alignment accuracy particularly when the matched residues are spatially apart (Figure 4A). For pairs of proteins with low sequence or structural homology, the SE algorithm produced correct alignments for up to 15% more residues than the next best scoring program that uses a dynamic programming algorithm. Another advantage is that its CPU usage grows linearly with the size of the protein (Figure 4B). Figures 4C and 4D show an example of the improved alignment. We expect that the SE algorithm will help improve the performance of not only SHEBA but all other programs that use the dynamic programming to produce the sequence alignment from a pair of superposed structures. This work is in progress.

Figure 4. **(A)** Comparison of the average alignment accuracy of SE, CHIMERA, LSQMAN, and SIPPL in SHEBA. CHIMERA (http://www.cgl.ucsf.edu/chimera/) and LSQMAN (http://xray.bmc.uu.se/usf/lsqman_man.html) are two programs that we could easily obtain that will produce an optimal sequence alignment given a pair of superposed structures. SIPPL is the name of the dynamic programming algorithm used in SHEBA for obtaining the optimal alignment from two superposed structures. It is so named because it was adapted from the original algorithm by Sippl's group (34). The 582 pairs of proteins collected from CDD were binned according to their RMSD based on CDD alignment and the average fraction of correctly aligned residues (f_{car}) of each program is shown. The number of pairs in each bin is 145, 260, 125, 43 and 9 for the RMSD ranges 0-1, 1-2, 2-3, 3-4, and 4-5 A, respectively.
(B) Comparison of CPU usage of SE and SIPPL routines in SHEBA. The CPU times to execute the SE algorithm and SIPPL routine in SHEBA were recorded. The 582 superimposed pairs are structurally similar according to CDD. The numbers of residues of the larger domain in each pair are shown in tan-colored dots.



(A)



(B)

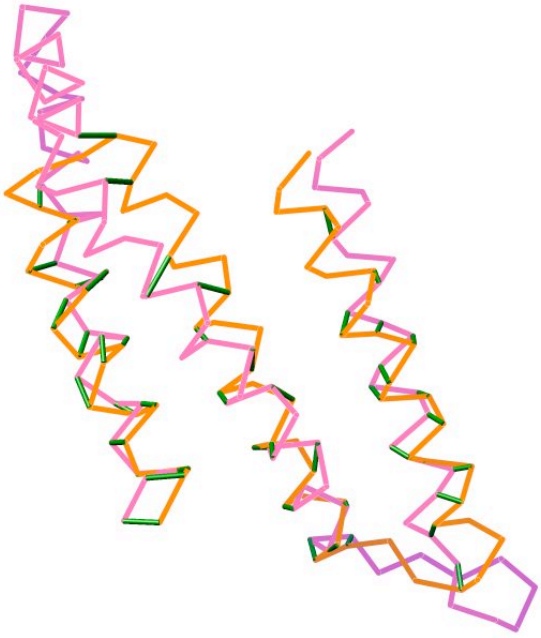
(C, D) Sequence alignments generated by SHEBA with SIPPL (C) and SE (D) routines for a pair of helical domains, 2FBW_Q (orange) and 1L0V_C(pink) in cd03493, after they were superimposed according to the CDD alignment. Pseudobonds in green indicate the residue pairs considered aligned by SHEBA with SIPPL (C) and with SE (D) routines. Unaligned regions do not have pseudobonds and are shown in lower case in the sequence alignments. The two proteins in the same CD family have three structurally homologous helices. However, the sequence alignment generated by SHEBA using dynamic programming (SIPPL) algorithm put many gaps in the second helix in the middle (f_{car} = 50%). SE does not apply gap penalty, yet generates accurate alignment without gaps (f_{car} = 100%).

SHEBA with Sippl (f_{car} = 50%)

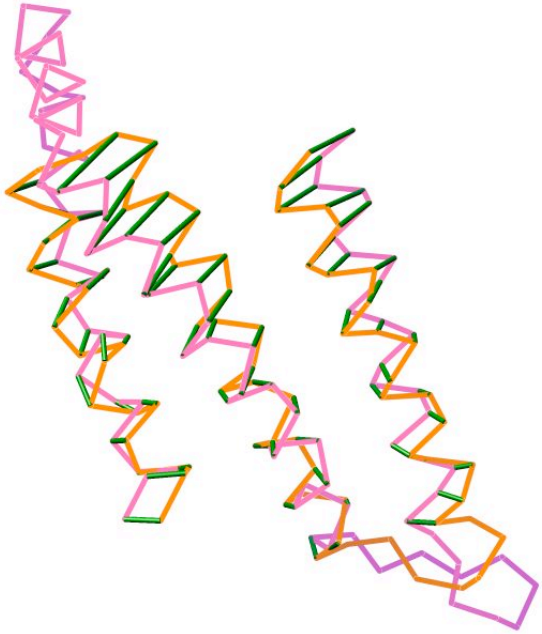
```
2FBW_Q   TSERAVSALL LGLLPaA--Y lyp-------
1L0V_C   MLREGTAVPA VWFSI-EliF glfalkngpe

2FBW_Q   --------Gp ----Avdy-- SlaA--AltL
1L0V_C   awagfvdfLq npviViinli T--LaaA--L

2FBW_Q   -HgH-WGlG- QVITDYvh-- ----GdtpI-
1L0V_C   lH-TkTW-Fe LAPKAAniiv kdekM---Gp

2FBW_Q   KVANTGLYVL SaItftGlcy --
1L0V_C   EPIIKSLWAV T-Vva-Tivi lf
```

SE (f_{car} = 100%)

```
TSERAVSALL LGLLPAAYLy p---------
MLREGTAVPA VWFSIELIFg lfalkngpea

---------- --GPAVDYSL AAALTLHGHW
wagfvdflqn pvIVIINLIT LAAALLHTKT

GLGQVITDyv hgdtp----I KVANTGLYVL
WFELAPKAan iivkdekmgP EPIIKSLWAV

SAITFTGLCY
TVVATIVILF
```



(C)



(D)

We would like to complement and complete this work by devising a new algorithm for finding the initial structural superposition more accurately with less failure. We are exploring the idea of using proper local structure hashing techniques, somewhat along the lines used by Alesker, Nussinov and Wolfson (35) and by Zhu and Weng (17), both for speed and for recognition and alignment of flexible structures.

(3) New protein domain parsing algorithm: The problem of parsing a protein structure into domains automatically is still an unsolved problem. We need domains because domain structures are the essential basic units for properly exploring the fold space or the relations between and among different protein structures. In the CASP6 experiment, we observed that three different automatic domain parsing programs (11) did not always agree with each other or with my manual parsing from visual inspection (unpublished results). Manual procedure is obviously unsuitable for processing a large number of structures because it is slow and tends to be subjective. We would like to re-visit this problem with the aim of producing an automatic procedure that follows the principles I used when I parsed protein structures manually for CASP6. The main principles used

then were (a) geometrical separation, (b) symmetry and (c) recurrence in other structures (11). The recurrence was used early by Holm and Sander (36), but the procedure was incorporated tightly into their PUU domain parsing procedure, which works like a protein unfolding simulator. My idea is to more directly follow the mental process that occurs when parsing the structures manually, except that whenever a decision needs to be made, it will be made not by subjective or ad hoc criteria but by statistical criteria established empirically from random cuts or by comparing unrelated structures. We plan to do this with Dr. Munson at CIT, who is a mathematician with expertise in statistics, and with Drs. Garnier and Gibrat of INRA, Jouy-en-Josas, France, who are experts in bioinformatics and algorithm development, respectively.

(4) Detection of repeating units within a protein structure: Many protein structures contain repeating units. Some examples are TIM barrel folds which typically have 8 β/α units; β-propeller folds, which have 4 to 8 antiparallel beta-sheet units; and superhelical solenoid structures made of the leucine-rich, ankyrin, or HEAT/ARM repeats (see below). Detecting such structures is important for at least two reasons. First, it helps in domain parsing procedure. There is some inherent ambiguity in defining domains for these proteins because it is not clear if a domain should consist of a single repeating unit or the whole protein (14). Also, some TIM barrels, for example, have noticeable separation between some groups of repeating units so that an automatic domain parsing procedure often splits the structure into two domains, whereas visual inspection invariably assigns one domain for the whole barrel. Therefore, automatic domain parsing procedure will be enormously simplified if these structures are recognized and treated separately. Secondly, many of these proteins have interesting properties. For example, TIM barrels are enzymes with widely varying substrate specificity while superhelical solenoid structures are involved in protein-protein interactions. In order to study the structure-sequence-function relations specific to each of these structural types, it is essential to be able to collect as many structures and proteins of one type as possible in order to see the generalities of an observation and to make statistical inferences.

Proteins with repeats are usually detected from observing repeating segments of sequence similarity within one protein sequence (37-39). However, sequence similarity between repeats is often low, in which case the protein is unrecognized and/or the repeats ill-defined. If the structure is known, a more sensitive procedure will be to detect structural, rather than sequence, similarity. Such structural symmetry-detecting routine is also needed for the domain parsing problem since symmetry was one of the considerations during manual parsing (14). There are only two reports that we know of on detecting proteins with repeating units using full structural information (40, 41). Both are from the same group of investigators and both detect repeats by analyzing the N x N inter-residue distance matrix, where N is the number of residues in the protein. However, they reported that the sensitivity of the method varied from a high of 88.7% for the Leucine rich repeat structures to a low of only about 60% for the TIM barrels (41). I would like to try a related, but different method, which involves structure alignments of the structure centered at different residues and analyzing the results using autocorrelation function(s). The advantage is that the method uses full structural information, not just the interresidue distances, and that one can use the best, among many, structure alignment programs.

(5) On-going protein structure modeling projects: We are currently working on two specific proteins:

(a) *Mesothelin* is a gene discovered in Pastan's laboratory and expressed in normal mesothelial cells and mesothelioma and ovarian and other cancers. An immunotoxin targeting this protein has been tested in phase I clinical trials. (See Dr. Pastan's Site Visit Report.) Mesothelin is produced as a part of the 69 kD precursor protein, which has a signal peptide at the N-terminus and a glycosylphosphatidylinositol (GPI) anchor peptide at the C-terminus. The furin cleavage of this precursor protein yields two proteins, the N-terminal megakaryocyte potentiating factor (MPF), which is secreted, and the C-terminal 329-residue mesothelin, which remains anchored to the membrane through GPI (42, 43). The mucin MUC16, which is expressed in ovarian cancer cells, strongly and specifically binds mesothelin through its N-linked oligosaccharides and it has been suggested that this interaction may facilitate peritoneal metastasis of ovarian tumors (44, 45). We started studying the possible structures of mesothelin to see if any of them resembled oligosaccharide-binding lectins. The PSI-BLAST search with mesothelin and mesothelin precursor sequences yielded only two non-mesothelin hits, stereocilin and otoancorin, which had low sequence similarity with mesothelin. Both are inner ear proteins anchored on the surface of hair cells and involved in mechanoreception of sound waves (46, 47). But structural information is not available for either protein. Several secondary structure prediction programs [psipred (48), profsec (49), DSC (50) and five other programs] gave the similar prediction of a series of small helical segments separated by short turns. This means that mesothelin does not have the typical lectin fold, which is a group of entirely beta-sheet structures (51). Three fold recognition and template-free prediction/modeling programs [INHUB (52), 3D-PSSM (30) and I-TASSER (53)] predicted an ARM/HEAT repeat α–α superhelical structure consistent with the predicted secondary structure. We have built a three-dimensional structural model of mesothelin based on these predictions and using the modeling software in Insight II. We are currently in the process of checking the internal consistency of the model. If the model appears reasonable, as is likely at present, this will be the first time, as far as we know, that an α–α superhelical structure is suggested to bind oligosaccharide.

We would of course love to build a model for the mesothelin-MUC16 complex. However, in the absence of the crystal structure of mesothelin or of the sugars of MUC16, it is unlikely that we will be able to build a useful model. We do, however, see a possible way to progress, by collaborating with the experimental group. One can mutate strategically placed residues on the surface of mesothelin, on the basis of the model, and see how the mutations affect the interaction with MUC16. One can obtain a fairly detailed model of the complex from a set of such mutation data, as we demonstrated some time ago with the galR repressosome in collaboration with Dr. Adhya of this laboratory (54).

(b) *The E. Coli Host Factor Q (HFQ)* is a pleiotropic regulator and believed to be an RNA chaperon protein in E. coli (55). It has been reported (56) that this protein also has an ATPase activity. The crystal structures of the protein from two bacterial species are

known (57, 58), but the ATP binding site is not obvious from these structures. We predicted the ATP binding site for this molecule. HFQ forms a ring of homo-hexamer. The structure is predominantly made of β-sheets, arranged in the shape of propeller blades, with one single helix per monomer. A well-known ATP-binding mode is through the Walker box, in which case the location of ATP is determined by its interaction with the P-loop, which usually connects a tip of a β-sheet with the N-terminal end of an α-helix (59). The ATP in HFQ cannot be in such a location since the lone helix is at the N-terminus of the protein and there is no P-loop. Dr. Adhya's group in this laboratory determined that mutating the residue Y25 nearly abolished the ATP binding activity. Y25 is on the face of the β-sheet, in the 'distal' side of the ring (60). It is known that ATP binds to a variety of protein structures, including a β-sheet (61). Therefore, we searched the structural database to see if there was an ATP binding protein in which the ATP is bound on the surface, rather than near the edge, of a β-sheet, and which also has a structure similar to that of HFQ around the Y25 residue. A search at the pdb site gave 213 entries with ATP-protein complexes, which together contained 742 SCOP domains from 81 different scop superfamilies. Only 35 of these superfamilies consisted of ATP-binding domains. When one randomly selected structure was examined from each of these 35 superfamilies using Insight II and the in-house graphics program GEMM, nine were found to have ATP on the face of a β-sheet. One of these was the d7at1b1 domain (SCOP d.58.2.1 family) of the B chain of 7at1 (aspartate carbamoyltransferase). When manually aligned using GEMM and Insight II, this domain superposed reasonably well with the HFQ monomer, although the two are clearly not homologous; 3 strands from one monomer and 2 strands from the adjacent monomer of HFQ align well with the β-sheet of d7at1b1 and a helix of d7at1b1 also superposes with the single helix of the HFQ monomer. The residue K94 of d7at1b1, which hydrogen bonds with ATP, aligns with K31 of hfq, which is near Y25 in the hexameric structure. The ATP molecule can be placed in the HFQ hexameric structure after only a small movement from its equivalent position in d7at1b1 when the two protein structures are superposed. After this position was found, we ran the binding site module calculation of Insight II. It gave two possible ATP binding sites, one around the rim of the central hole of the hexameric ring and another at the site we identified from the d7at1b1 structure. This work is nearly completed and ready to be written up.

(6) Study of superhelical structures with repeat motif: Three of the cancer gene products we study, POTE, CAPC, and mesothelin, are predicted to have structures that are made of repeating units arranged in a superhelical manner: POTE with the ankyrin repeats (62), CAPC with the leucine-rich repeats (63), and mesothelin with the probable HEAT/ARM type repeats. These repeat structures occur frequently in all types of organisms and have been extensively studied (38, 64, 65). Most, if not all, of these structures are involved in protein-protein interactions (66). The defining characteristic of these structures is that each unit interacts only with its immediate neighbor repeating units, without longer range interactions. This gives the structure flexibility, which may be required in order to make good interaction with a variety of structures. As a result, the repeating units have similar structures but their relative position and orientation often vary so that overall shape can deviate noticeably from an ideal superhelical symmetry and from one structure to another. We would like to study how these structures interact

with other proteins and small ligands, what type of residues in which structural context are involved in the complex formation, and how the conformation of these structures change upon such interactions. We will begin by collecting single and complex structures involving these superhelical structures and by classifying them. Sequence-specific features and conformational changes will be noted for each type. When needed, the changes in flexibility of the molecule upon mutation will be studied using conventional molecular dynamics programs.

PROJECT 2: HYDROPHOBICITY

Background

The phenomenon of hydrophobicity refers to the low solubility of non-polar molecules in water and, hence, their tendency to group together and separate out from water. The hydrophobic effect is believed to be one of the main forces that determine the structure, stability, and interaction of protein and other biologically important molecules (67). Rather surprisingly, there is as yet no consensus on what causes this phenomenon and various other associated thermodynamic behaviors nor on the true magnitude of the hydrophobic effect. I study this phenomenon by means of the statistical thermodynamics. Some years ago, I proposed that the primary reason that a non-polar molecule avoids water is the small size, not the hydrogen bonding capability, of water molecules (68-70). This theory is now slowly being adopted (71, 72). I and my colleague, Dr. Giuseppe Graziano of University of Sannio, Italy, also modified Muller's simple two-state model of water of hydration (73) to put it on a firmer statistical mechanical footing (74). This modified Muller's model explains the large entropy and heat capacity changes upon hydration without assuming "iceberg"-like structure of water around the non-polar solute molecule. The model has since been used by other researchers (75) to study the hydrophobic hydration. My work in this area is now restricted to collaboration with Dr. Graziano, who is highly productive in this area on his own.

Specific Research Aims

My aim is to obtain molecular level understanding of various aspects of this complex phenomenon.

Accomplishments

In the past four years, Dr. Graziano and I produced two significant works:

(1) We showed that the scaled particle theory can be used to explain the phenomenon of the entropy convergence (76). The phenomenon of entropy convergence refers to the fact that the entropies of hydration of different solute molecules, when plotted as a function of temperature, converge to the same or similar value at certain temperature, usually near the boiling point of water. I proved years ago that precise convergence will occur if the hydration entropy varied linearly with some property of the solute molecules, for example the surface area (77). The proof was based on the mathematical property of a

bilinear function and had little to do with statistical mechanical physics of hydration. The current work is important because it proves that the same phenomenon can be explained from a statistical mechanical theory with no other assumptions. Since the equations from the statistical mechanical theory are not linear, the convergence is obtained from locally linear behavior and there is a significant de-focusing, or lack of precise convergence, as is the case for real experimental data.

(2) We proved that the structure of water must decrease in the hydration shell in the modified Muller's model (78). A big question in the hydrophobic phenomenon is the state of water in the hydration shell of a non-polar molecule. When a non-polar molecule is inserted in water, the entropy of the system decreases markedly. Since at room temperature, the enthalpy change is nearly zero, the large reduction in entropy is the reason that non-polar molecules do not dissolve in water. It has been traditional to suppose that this large negative change in entropy means that there is more structure formation in the hydration shell. However, I always considered this explanation to be erroneous since introduction of a non-polar molecule should disrupt, not enhance, the water structure. I found another, more plausible source of entropy reduction in the small size of water molecules. In the current work, Dr. Graziano and I mathematically proved that, within the framework of the modified Muller's model, the water structure must break compared to the bulk if the heat capacity change were to decrease with temperature as observed experimentally.

Current Research and Future Plans

I am not currently working on any project on hydrophobicity. My future plan on this project is opportunistic. I will work on this topic if and when I find a significant solvable problem.

PROJECT 3: IMMUNOTOXIN AND GENE DISCOVERY

Background

Dr. Ira Pastan's group is developing a novel type of anti-cancer agents called immunotoxin. These are man-made molecules made by fusing part of a bacterial toxin (pseudomonas exotoxin) to the Fv portion of mouse antibodies against cancer-specific targets. We collaborate with this group to provide molecular modeling and other computational support. There is also a need to find more specific targets in order to reduce the side effect and more targets in order to treat different types of cancer. We have set up a procedure for searching through the expressed sequence (EST and mRNA) databases to discover genes that appear to be expressed specifically in breast or prostate and in cancer cells, but not in other essential tissues. The products of such genes can be used as a new target for the immunotoxin or for cancer vaccine, diagnosis, and imaging. We periodically update databases used in this procedure and generate new list of candidate genes. All the studies on this project are in collaboration with Dr. Pastan's group.

Specific Research Aims

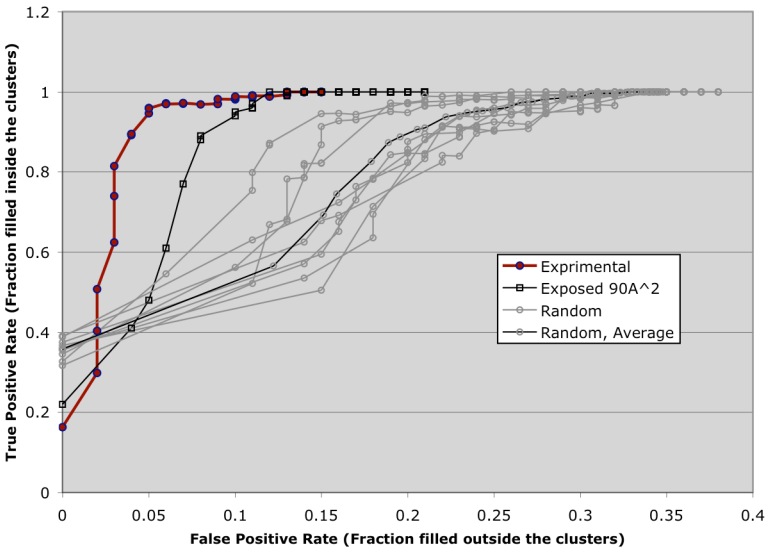
Sub-project 1. To help find major epitopes of the immunotoxins. Pastan's group generated 60 mouse monoclonal antibodies and quantitatively determined all-against-all pair-wise competition in binding to the immunotoxin. On the basis of this data, they could group the antibodies into 7 to 13 groups and interpreted the result as indicating that there are 7 to 13 major epitopes against which all the monoclonal antibodies respond (79). Our aim was to find a mathematical way of validating this interpretation.
Sub-project 2. To discover genes that are expressed specifically in cancer and normal but non-essential tissues and to use bioinformatics techniques to gather as much information as possible on the possible function of the product.

Accomplishments

Sub-project 1. We devised a simple model for random competitive binding on the surface of the immunotoxin molecule and a new ROC (Receiver-Operator Characteristic) curve-based method to measure the quality of clustering of the antibodies. In this model we draw a roughly equal size footprint on the surface of the immunotoxin to represent an antibody binding to an epitope. A set of footprints are placed randomly on the surface and two antibodies are considered to compete if their corresponding footprints overlap. This model produces N x N competition matrix where N is the number of antibodies. Given a set of N x N competition data, one can optimally group the antibodies into a given number, C, of clusters. The quality of the clustering can be measured by the number of competing pairs within each cluster (true positives) and the number of competing pairs that belong to two different clusters (false positives). This is essentially the NDO scoring scheme that we used for protein domain parsing (see above), but simplified to these straightforward clusters. (Protein domain parsing is more complicated because of the need to handle linkers between domains and the segmented domains.) A ROC curve is obtained by plotting the fraction of true positives against that of false positives as the number of clusters, C, is varied from 1 to N. For a set of perfectly discrete epitopes, the ROC curve coincides with the y-axis for C = 1 to M, where M is the number of discrete epitopes, then it coincides with the y = 1 horizontal line from C = M to N. For a random selection of epitopes, the fraction of the true and false positives will be roughly equal and the ROC curve tracks the diagonal. We generated 10 sets of 60 random epitopes each, to which the 60 antibodies bind (and leave foot prints), and observed that their artificial competition data generate ROC curves that were much different from the ROC curve obtained from the real experimental competition data (Figure 5). The simple simulation data, therefore, supports the notion that the pattern of the experimentally observed competitive binding could not have been obtained if the antibodies bound to a random set of epitopes and that the pattern was indeed consistent with there being a discrete set of epitopes on the surface of the immunotoxins (80).

Figure 5. ROC curves for the evaluation of the clustering of the epitopes of PE38. The experimental competition data among 60 mouse monoclonal antibodies are represented by black circles connected by red lines. Other data on the graph (grey circles and black squares) are for the artificial data generated from overlaps of model antibody footprints.

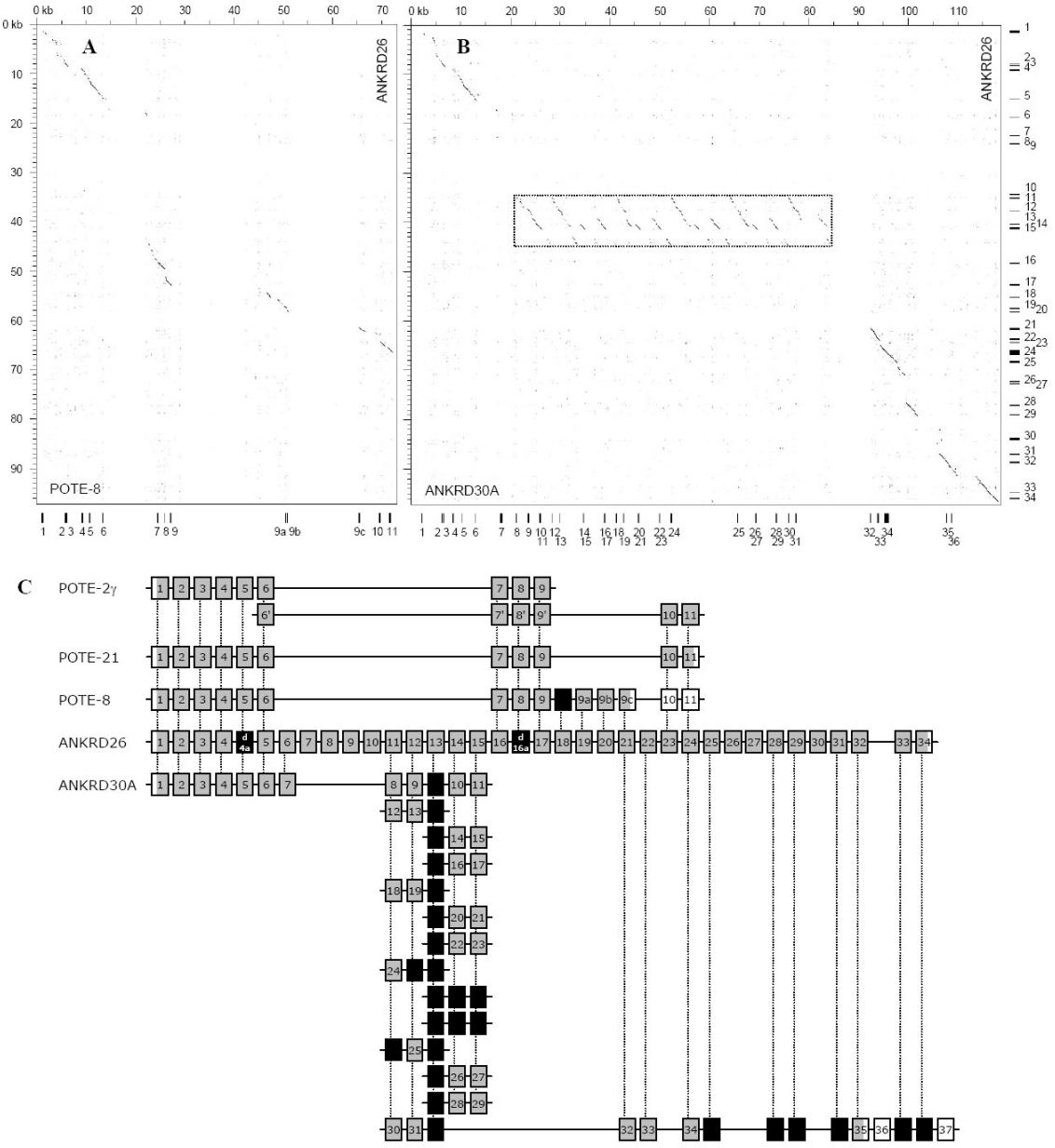
These latter were generated by randomly selecting 60 residues from all exposed residues (grey circles with grey lines) or from those that had more than 90 A^2 of exposed surface area (black squares with black lines) on the surface of PE38, drawing a circle around each to form the initial footprint, and then expanding it radially until the footprint included 60 or more atoms. The elements of the competition matrix for this set of artificial epitopes were set to 1 if the corresponding pair of footprints overlap and 0 if not. The datasets with the randomly selected exposed residues (grey circles) were generated 10 times independently by repeating the same procedure. The grey circles with black lines indicate the average. Both the experimental and the artificial competition matrices were clustered using the Ward's hierarchical clustering method for each given number of clusters. It is clear that randomly distributed epitopes generate ROC curves (grey circles) that lie close to the diagonal of the plot and which are clearly different from that of the experimental data. When the epitopes were selected randomly among a restricted set of highly exposed residues only, the ROC curve (black squares) more closely resembles the experimental curve.



Sub-project 2. During this reporting period, we reported the discovery of four additional genes, *PRAC2* (81), *NGEP*, *POTE*, and *CAPC*, which are specific to prostate and/or breast and various cancers. NGEP is a member of the TMEM16 family of proteins of unknown function and predicted to be an integral membrane protein, with 8 trans-membrane domains (82, 83). The probable location of the protein on the cell surface and the high specificity of expression in prostate and prostate cancer make this a promising candidate for a new immunotoxin target. *POTE* is a primate-specific family of genes (84). There are 13 paralogs in humans, scattered across 8 chromosomes. All paralogs are in the pericentromeric region, except those on chromosome 2, which are in an old, degenerating pericentromeric region (85). Their expression pattern is restricted to the prostate, ovary, testis, embryonic stem cells and in many cancers (86). Since it contains ankyrin repeats, spectrin-like coiled coil region and actin in some paralogs, we expect it to be located at the cytoplasmic aspect of the membrane, connecting it to the cytoskeleton (84, 87). We also found an ancient gene, ANKRD26, which appears to be the ancestor of the POTE gene family (85) (Figure 6). The function of ANKRD26 is unknown, but it must be a critically important gene, since it is expressed at a low level in many different cell types and is present, and its sequence highly conserved, in organisms from sea urchin to man. Recently, Pastan's group found that a disruption of this gene by a gene trap

technique causes extreme obesity and increase in body size in homozygous mice. (See Dr. Pastan's Site Visit Report.) CAPC is made of leucine-rich repeats, one putative transmembrane domain, and a short cytoplasmic tail at the C-terminus (63). It is expressed in breast, prostate, and salivary gland as well as in many cancers. Function is unknown. A phylogenetic analysis of CAPC orthologs from mammals shows that the putative cytoplasmic tail may be subject to rapid evolution. Interestingly, primates have what appears to be a primate-specific stretch of highly proline-rich sequence (Figure 7).

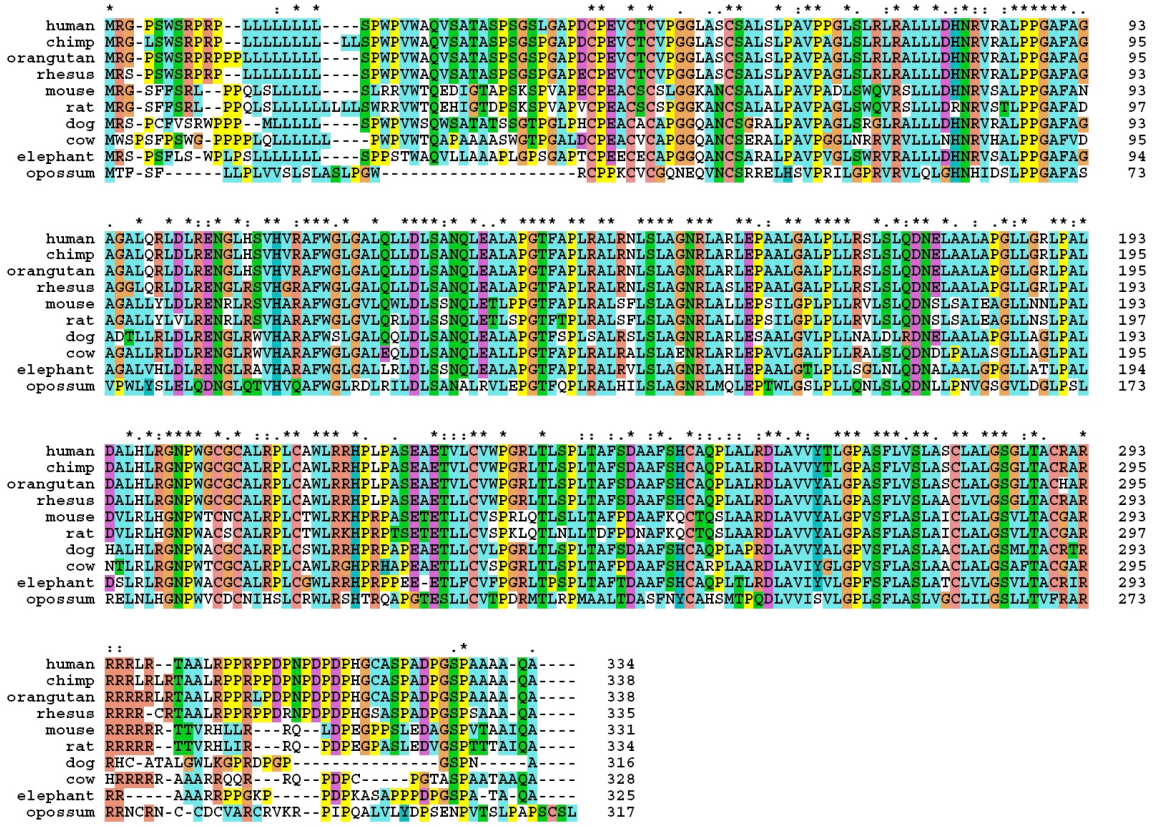
Figure 6.



Dot-plots show conserved genomic fragments **(A)** between *POTE*-8 and *ANKRD26* and **(B)** between *ANKRD30A* and *ANKRD26*. The ruler is in kb units both at the top and at the

left. Exons are marked as lines with numbers at the bottom and at the right. Duplicated segments are marked by a dotted box.
(C) Schematic representation of the genomic organization of the exons of *POTE* and its related genes. Homologous exons of *POTE*-2γ, *POTE*-21, *POTE*-8, *ANKRD26*, and *ANKRD30A* are aligned along dotted lines. In the cases of *POTE*-2γ and *ANKRD30A*, the genomic region continues to the next line(s). The open reading frames are in gray. Black boxes are degenerated exons which are not represented in the mRNA sequences but exist in the corresponding genomic region. The horizontal lines are drawn between the exon boxes of each gene – their lengths do not indicate the lengths of the introns.

Figure 7. Multiple alignment of CAPC protein sequences from mammalian species. The symbols above the sequences indicate the conservation level: *, :, and . for fully, highly, and moderately conserved sites, respectively. The LRRs and transmembrane domains in the middle portion of the sequences are highly conserved. The N-terminal signal peptide region and the C-terminal region after the transmembrane domain are less well conserved. Note the high proline content in the C-terminal region of the primate sequences.



Current Research and Future Plans

(1) Mathematical model of the immunotoxin delivery process: The toxin is potent; it has been estimated that just a few molecules inside the cell can kill the whole cell. However, the effective dose found to reduce the tumor size of mouse xenograph model

corresponds to more than several hundred molecules per tumor cell. (See Dr. Pastan's Site Visit Report.) We are working on making a mathematical model of the delivery process to provide a quantitative understanding of the process and to identify the sources of waste and to help determine the dosing method and other ways to make the delivery process more efficient.

The mathematical model consists of a set of differential equations that represent the rates of various processes that include the translocation of the immunotoxin (IT) from blood vessel into the tumor tissue, diffusion through the intercellular space of the tumor tissue, non-specific decay and clearance from the tumor tissue, uptake by the tumor cells, endocytosis and transport through the cell interior into the cytoplasm, decay during this process, cell killing, and the tumor volume change (growth or shrinkage). We have built an initial model, determined a reasonable set of values for the many parameters of the model, and fairly accurately reproduced the experimentally observed tumor volume change upon IT administration on mouse xenograph tumor models. Unfortunately, the progress on this project has been slow beyond this initial stage because of a personnel change during this reporting period. I plan to pursue this project vigorously in the coming years. We will complete the mathematical modeling and provide precise accounting of the IT lost during the delivery process. In addition, new experimental data show that a significant concentration of shed antigen is present in the intercellular space of the tumor tissue. (See Dr. Pastan's Site Visit Report.) This will substantially change the concentration of the IT actually delivered inside the tumor cell. We will modify our model to include the effect of the shed antigen. Once the model is built and tested, we will use and study the model to discover the bottleneck(s) in the delivery process and search for ways to remove the bottleneck(s).

(2) Gene discovery: We will not attempt to discover more new genes for immunotoxin targets using the expressed sequence database, mainly because we now have a large number of genes to study. However, we will continue to collect information on the functional and structural features of the proteins encoded by these genes and work with the experimental group to design experiments that will shed light on the biological function of the gene. (See above for the structural modeling of mesothelin and the study of the structures with repeating units.)

PROJECT 4: EXPLORATION OF THE HUMAN GENOME

Background

During our search for genes that are specific to prostate, breast, and cancer, we encountered transcript sequences that appear to have been derived from two distinct genes. Most of these are cloning artifacts, introduced by inadvertent joining of two different cDNA sequences. However, some of these must be from real chimeric genes, which are produced when two distinct genes are fused together by chromosomal aberrations known to occur in all cancer cells. If a method can be devised for identifying such transcripts, it would provide a window for a high resolution view of the fusion

points of some of the chromosomal aberration events and possibly for identifying genes that are responsible for generation and/or maintenance of cancer.

We have also found that some of the genes we discovered appear to be inactive only in humans because the translation is terminated by a premature stop codon. Since these are genes that suffered a recent mutation with a rather severe consequence in terms of the activity of the particular gene, they may give a clue to the development of some human-specific traits and possibly to diseases that humans are prone to suffer.

Human genome is made of three billion bases and difficult to explore. Above studies give us a handle for studying human genes in both the healthy and diseased states.

Specific Research Aims

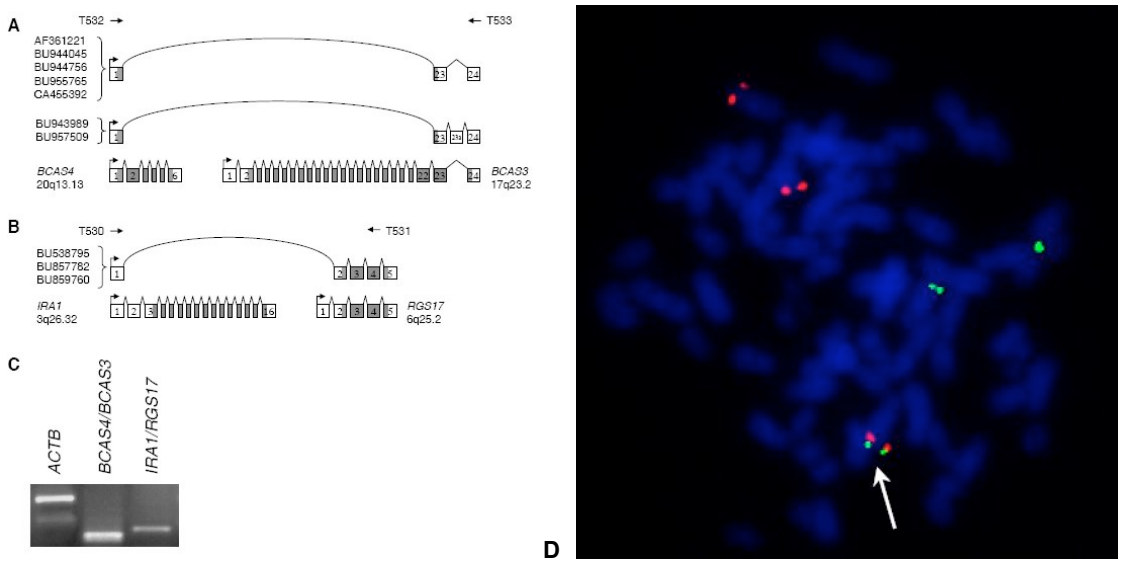
Sub-project 1. To systematically identify active chimeric fusion genes produced by chromosomal aberrations.
Sub-project 2. To find genes that have been inactivated or otherwise altered specifically in humans by frameshift, nonsense, and exon deletion mutations.

Accomplishments

Sub-project 1. We devised an algorithm that will distinguish chimeric transcripts that are derived from true fusion genes and those that are cloning artifacts. The algorithm is based on the principle that the fusion point for the cloning artifact will usually be in the middle of an exon whereas that from a real chimeric gene will be at an exon boundary since the chromosomal break is much more likely to happen in a long intron than in a relatively short exon. Searching through the mRNA and EST (Expressed Sequence Tag) databases systematically using this algorithm, and then screening further manually, we could identify 237 fusion cases in different tissue types that produced 314 transcript sequences (88). About a quarter (60) of these cases were known cases already described in the literature. The remainder represents new cases that have not been described before. An example is shown in Figure 8. A surprising finding is that 92 of the cases identified are from unmarked, presumably normal tissues.

Figure 8. Schematic representation and RT-PCR detection of *BCAS4/BCAS3* and *IRA1/RGS17* fusions in MCF7 cells. The *BCAS4/BCAS3* fusion **(A)** had been described in the literature; the *IRA1/RGS17* fusion **(B)** was newly discovered in this study. Boxes represent the exons and broken lines the introns. Fusion events are indicated by the arcs. Arrows indicate the transcription start sites. Exons are numbered from the 5' to the 3' direction as they occur in the original gene. Two *BCAS4/BCAS3* fusion transcripts, BU943989 and BU957509, have an additional exon between *BCAS3* gene exons 23 and 24, which is designated as 23a. Primers for the RT-PCR reaction are indicated (T530, T531, T532, and T533). ORFs are marked with grey boxes. **(C)** The fusion gene transcripts for the *BCAS4/BCAS3* and the *IRA1/RGS17* fusions were detected in MCF7 cells. The β actin (*ACTB*) was used as the positive control. The product sizes of *ACTB*, *BCAS4/BCAS3*, and *IRA1/RGS17* are 600, 328, and 367 bp, respectively. **(D)** Detection

of the 3;6 translocation in MCF7 cells by FISH of metaphase chromosomes. A representative result of the FISH experiment is presented. The *IRA1* gene (red) and the *RGS17* gene (green) are on the chromosomes 3 and 6, respectively, each of which exists in two copies of sister chromatid pairs. In addition, one can see another sister chromatid pair which harbors both genes in the same chromosome (white arrow).



Sub-project 2. To detect human specific frameshift, nonsense, and exon deletion mutations, we compared the chimpanzee and human genome sequences to detect differences that could potentially arise from these different mutations, used one or more non-human, non-chimpanzee sequence as an outgroup to determine if the mutation occurred in the human lineage, and then manually verified the candidate cases. The number of genes that harbor the mutations, after a rather stringent manual selection, were 9, 9, and 6, respectively, for the frameshift, nonsense, and exon deletion mutations (89-91). In 7 of these 24 mutation cases, the gene appears to have been totally inactivated in the human lineage. Interestingly, 6 of the 9 nonsense mutations were polymorphic in human population, suggesting that the mutations occurred rather recently and have not yet been fixed in the entire human population. Some of the interesting cases found are:

NPPA: The human-specific form has a nonsense mutation near the 3'-end of the coding sequence, which deletes the terminal two arginine residues in the protein product (90). The gene is polymorphic in human; 17% of the human chromosomes carry the original chimpanzee form. It has been reported that individuals homozygous for the ancestral form are associated with a significantly increased risk of ischemic stroke recurrence (92).

MOXD2: The human-specific form lost two terminal exons, which include 3' UTR and poly (A) signal as well as nearly a quarter of the 618 residue protein coding region, including the C-terminal GPI anchor residues (91). The gene bears a homology with dopamine beta hydroxylase (DBH), is highly conserved in animal species, and in mouse is highly expressed in medial olfactory epithelium.

S100A15A: The human form lacks the first of the two coding exons in the chimpanzee gene, which includes the start codon (91). The S100 proteins are calcium-binding proteins. The mouse ortholog *s100a15* was detected in differentiating cells of the hair follicles and cornified layer during skin maturation (93, 94). The gene has also been reported to be expressed in mammary gland and upregulated during mammary tumorigenesis (95). Humans have a functioning paralog, S100A7a, which is also known as S100A15 and wrongly considered as the human ortholog of mouse s100a15 (93, 94). The sequence of this paralog is sufficiently different from that of S100A15A that the functions of these two genes may have been distinct when the latter was functioning (91).

Current Research and Future Plans

We are currently looking for genes that harbor human-specific exon insertion mutations using a similar technique.

REFERENCES

1. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379-400 (1971).
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
3. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
4. Henikoff, S. & Henikoff, J. G. Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49-61 (1993).
5. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
6. Jung, J. & Lee, B. Protein structure alignment using environmental profiles. *Protein Eng* **13**, 535-543 (2000).
7. Brenner, S. E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* **28**, 254-256 (2000).
8. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
9. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251 (2006).
10. Goonesekere, N. C. & Lee, B. Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. *Nucleic Acids Res* **32**, 2838-2843 (2004).
11. Tress, M., Tai, C. H., Wang, G., Ezkurdia, I., Lopez, G., Valencia, A., Lee, B. & Dunbrack, R. L., Jr. Domain definition and target classification for CASP6. *Proteins* **61 Suppl 7**, 8-18 (2005).

12. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370-3374 (2003).
13. Vincent, J. J., Tai, C. H., Sathyanarayana, B. K. & Lee, B. Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins* **61 Suppl 7**, 67-83 (2005).
14. Tai, C. H., Lee, W. J., Vincent, J. J. & Lee, B. Evaluation of domain prediction in CASP6. *Proteins* **61 Suppl 7**, 183-192 (2005).
15. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-747 (1998).
16. Holm, L. & Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics* **16**, 566-567 (2000).
17. Zhu, J. & Weng, Z. FAST: a novel protein structure alignment algorithm. *Proteins* **58**, 618-627 (2005).
18. Shapiro, J. & Brutlag, D. FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res* **32**, W536-541 (2004).
19. Kawabata, T. MATRAS: A program for protein 3D structure comparison. *Nucleic Acids Res* **31**, 3367-3369 (2003).
20. Madej, T., Gibrat, J. F. & Bryant, S. H. Threading a database of protein cores. *Proteins* **23**, 356-369 (1995).
21. Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Curr Opin Struct Biol* **6**, 377-385 (1996).
22. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Yamashita, R. A., Yin, J. J., Zhang, D. & Bryant, S. H. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* **33**, D192-196 (2005).
23. Pearson, W. R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635-650 (1991).
24. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-229 (2004).
25. Sam, V., Tai, C. H., Garnier, J., Gibrat, J. F., Lee, B. & Munson, P. J. ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. *BMC Bioinformatics* **7**, 206 (2006).
26. del Solar, G., Giraldo, R., Ruiz-Echevarria, M. J., Espinosa, M. & Diaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62**, 434-464 (1998).
27. Komori, H., Matsunaga, F., Higuchi, Y., Ishiai, M., Wada, C. & Miki, K. Crystal structure of a prokaryotic replication initiator protein bound to DNA at 2.6 A resolution. *Embo J* **18**, 4597-4607 (1999).
28. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280 (2002).
29. Fischer, D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac Symp Biocomput*, 119-130 (2000).

30. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* **299**, 499-520 (2000).
31. Sharma, S., Sathyanarayana, B. K., Bird, J. G., Hoskins, J. R., Lee, B. & Wickner, S. Plasmid P1 RepA is homologous to the F plasmid RepE class of initiators. *J Biol Chem* **279**, 6027-6034 (2004).
32. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* **A32**, 922-923 (1976).
33. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* **A34**, 827-828 (1978).
34. Feng, Z. K. & Sippl, M. J. Optimum superimposition of protein structures: ambiguities and implications. *Fold Des* **1**, 123-132 (1996).
35. Alesker, V., Nussinov, R. & Wolfson, H. J. Detection of non-topological motifs in protein structures. *Protein Eng* **9**, 1103-1119 (1996).
36. Holm, L. & Sander, C. Dictionary of recurrent domains in protein structures. *Proteins* **33**, 88-96 (1998).
37. Heringa, J. Detection of internal repeats: how common are they? *Curr Opin Struct Biol* **8**, 338-345 (1998).
38. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. *J Struct Biol* **134**, 117-131 (2001).
39. Soding, J., Remmert, M. & Biegert, A. HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res* **34**, W137-142 (2006).
40. Taylor, W. R., Heringa, J., Baud, F. & Flores, T. P. A Fourier analysis of symmetry in protein structure. *Protein Eng* **15**, 79-89 (2002).
41. Murray, K. B., Taylor, W. R. & Thornton, J. M. Toward the detection and validation of repeats in protein structure. *Proteins* **57**, 365-380 (2004).
42. Kojima, T., Oh-eda, M., Hattori, K., Taniguchi, Y., Tamura, M., Ochi, N. & Yamaguchi, N. Molecular cloning and expression of megakaryocyte potentiating factor cDNA. *J Biol Chem* **270**, 21984-21990 (1995).
43. Chang, K. & Pastan, I. Molecular cloning of mesothelin, a differentiation antigen present on mesothelium, mesotheliomas, and ovarian cancers. *Proc Natl Acad Sci U S A* **93**, 136-140 (1996).
44. Rump, A., Morikawa, Y., Tanaka, M., Minami, S., Umesaki, N., Takeuchi, M. & Miyajima, A. Binding of ovarian cancer antigen CA125/MUC16 to mesothelin mediates cell adhesion. *J Biol Chem* **279**, 9190-9198 (2004).
45. Gubbels, J. A., Belisle, J., Onda, M., Rancourt, C., Migneault, M., Ho, M., Bera, T. K., Connor, J., Sathyanarayana, B. K., Lee, B., Pastan, I. & Patankar, M. S. Mesothelin-MUC16 binding is a high affinity, N-glycan dependent interaction that facilitates peritoneal metastasis of ovarian tumors. *Mol Cancer* **5**, 50 (2006).
46. Verpy, E., Masmoudi, S., Zwaenepoel, I., Leibovici, M., Hutchin, T. P., Del Castillo, I., Nouaille, S., Blanchard, S., Laine, S., Popot, J. L., Moreno, F., Mueller, R. F. & Petit, C. Mutations in a new gene encoding a protein of the hair bundle cause non-syndromic deafness at the DFNB16 locus. *Nat Genet* **29**, 345-349 (2001).
47. Zwaenepoel, I., Mustapha, M., Leibovici, M., Verpy, E., Goodyear, R., Liu, X. Z., Nouaille, S., Nance, W. E., Kanaan, M., Avraham, K. B., Tekaia, F., Loiselet, J., Lathrop, M., Richardson, G. & Petit, C. Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying

acellular gels, is defective in autosomal recessive deafness DFNB22. *Proc Natl Acad Sci U S A* **99**, 6240-6245 (2002).
48. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202 (1999).
49. Rost, B., Yachdav, G. & Liu, J. The PredictProtein server. *Nucleic Acids Res* **32**, W321-326 (2004).
50. King, R. D. & Sternberg, M. J. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* **5**, 2298-2310 (1996).
51. Rini, J. M. Lectin structure. *Annu Rev Biophys Biomol Struct* **24**, 551-577 (1995).
52. Fischer, D. & Eisenberg, D. Protein fold recognition using sequence-derived predictions. *Protein Sci* **5**, 947-955 (1996).
53. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17 (2007).
54. Geanacopoulos, M., Vasmatzis, G., Lewis, D. E., Roy, S., Lee, B. & Adhya, S. GalR mutants defective in repressosome formation. *Genes Dev* **13**, 1251-1262 (1999).
55. Brennan, R. G. & Link, T. M. Hfq structure, function and ligand binding. *Curr Opin Microbiol* **10**, 125-133 (2007).
56. Sukhodolets, M. V. & Garges, S. Interaction of Escherichia coli RNA polymerase with the ribosomal protein S1 and the Sm-like ATPase Hfq. *Biochemistry* **42**, 8022-8034 (2003).
57. Schumacher, M. A., Pearson, R. F., Moller, T., Valentin-Hansen, P. & Brennan, R. G. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. *Embo J* **21**, 3546-3556 (2002).
58. Sauter, C., Basquin, J. & Suck, D. Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli. *Nucleic Acids Res* **31**, 4091-4098 (2003).
59. Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci* **15**, 430-434 (1990).
60. Mikulecky, P. J., Kaw, M. K., Brescia, C. C., Takach, J. C., Sledjeski, D. D. & Feig, A. L. Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nat Struct Mol Biol* **11**, 1206-1214 (2004).
61. Denessiouk, K. A. & Johnson, M. S. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins* **38**, 310-326 (2000).
62. Bera, T. K., Zimonjic, D. B., Popescu, N. C., Sathyanarayana, B. K., Kumar, V., Lee, B. & Pastan, I. POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer. *Proc Natl Acad Sci U S A* **99**, 16975-16980 (2002).
63. Egland, K. A., Liu, X. F., Squires, S., Nagata, S., Man, Y. G., Bera, T. K., Onda, M., Vincent, J. J., Strausberg, R. L., Lee, B. & Pastan, I. High expression of a cytokeratin-associated protein in many cancers. *Proc Natl Acad Sci U S A* **103**, 5929-5934 (2006).
64. Enkhbayar, P., Kamiya, M., Osaki, M., Matsumoto, T. & Matsushima, N. Structural principles of leucine-rich repeat (LRR) proteins. *Proteins* **54**, 394-403 (2004).

65. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. Y. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci* **13**, 1435-1448 (2004).
66. Kobe, B. & Kajava, A. V. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem Sci* **25**, 509-515 (2000).
67. Baldwin, R. L. Energetics of protein folding. *J Mol Biol* **371**, 283-301 (2007).
68. Lee, B. in Mathematics and Computers in Biomedical Applications (eds. Eisenfeld, J. & DeLisi, C.) 3-11 (Elsevier, North-Holland, 1985).
69. Lee, B. The physical origin of the low solubility of nonpolar solutes in water. *Biopolymers* **24**, 813-823 (1985).
70. Lee, B. Solvent reorganization contribution to the transfer thermodynamics of small nonpolar molecules. *Biopolymers* **31**, 993-1008 (1991).
71. Blokzijl, W. & Engberts, J. B. F. N. Hydrophobic effect. Opinions and facts. *Angewandte Chemie International Edition in English* **32**, 1545-1579 (1993).
72. Buchanan, P., Aldiwan, N., Soper, A. K., Creek, J. L. & Koh, C. A. Decreased structure on dissolving methane in water. *Chemical Physics Letters* **415**, 89-93 (2005).
73. Muller, N. Search for a realistic view of hydrphobic effects. *Accounts of Chemical Research* **23**, 23-28 (1990).
74. Lee, B. & Graziano, G. A two-state model of hydrophobic hydration that produces compensating enthalpy and entropuy changes. *Journal of the American Chemical Society* **118**, 5163-5168 (1996).
75. Silverstein, K. A. T., Haymet, A. D. J. & Dill, K. A. The strength of hydrogen bonds in liquid water around nonpolar solutes. *Journal of the American Chemical Society* **122**, 8037-8041 (2000).
76. Graziano, G. & Lee, B. Entropy convergence in hydrophobic hydration: a scaled particle theory analysis. *Biophys Chem* **105**, 241-250 (2003).
77. Lee, B. Isoenthalpic and isoentropic temperatures and the thermodynamics of protein denaturation. *The Proceedings of the National Academy of Sciences USA* **88**, 5154-5158 (1991).
78. Graziano, G. & Lee, B. On the intactness of hydrogen bonds around nonpolar solutes dissolved in water. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys* **109**, 8103-8107 (2005).
79. Onda, M., Nagata, S., FitzGerald, D. J., Beers, R., Fisher, R. J., Vincent, J. J., Lee, B., Nakamura, M., Hwang, J., Kreitman, R. J., Hassan, R. & Pastan, I. Characterization of the B cell epitopes associated with a truncated form of Pseudomonas exotoxin (PE38) used to make immunotoxins for the treatment of cancer patients. *J Immunol* **177**, 8822-8834 (2006).
80. Onda, M., Nagata, S., Ho, M., Bera, T. K., Hassan, R., Alexander, R. H. & Pastan, I. Megakaryocyte potentiation factor cleaved from mesothelin precursor is a useful tumor marker in the serum of patients with mesothelioma. *Clin Cancer Res* **12**, 4225-4231 (2006).
81. Olsson, P., Motegi, A., Bera, T. K., Lee, B. & Pastan, I. PRAC2: a new gene expressed in human prostate and prostate cancer. *Prostate* **56**, 123-130 (2003).
82. Bera, T. K., Das, S., Maeda, H., Beers, R., Wolfgang, C. D., Kumar, V., Hahn, Y., Lee, B. & Pastan, I. NGEP, a gene encoding a membrane protein detected only in

prostate cancer and normal prostate. *Proc Natl Acad Sci U S A* **101**, 3059-3064 (2004).
83. Das, S., Hahn, Y., Nagata, S., Willingham, M. C., Bera, T. K., Lee, B. & Pastan, I. NGEP, a prostate-specific plasma membrane protein that promotes the association of LNCaP cells. *Cancer Res* **67**, 1594-1601 (2007).
84. Bera, T. K., Huynh, N., Maeda, H., Sathyanarayana, B. K., Lee, B. & Pastan, I. Five POTE paralogs and their splice variants are expressed in human prostate and encode proteins of different lengths. *Gene* **337**, 45-53 (2004).
85. Hahn, Y., Bera, T. K., Pastan, I. H. & Lee, B. Duplication and extensive remodeling shaped POTE family genes encoding proteins containing ankyrin repeat and coiled coil domains. *Gene* **366**, 238-245 (2006).
86. Bera, T. K., Saint Fleur, A., Lee, Y., Kydd, A., Hahn, Y., Popescu, N. C., Zimonjic, D. B., Lee, B. & Pastan, I. POTE paralogs are induced and differentially expressed in many cancers. *Cancer Res* **66**, 52-56 (2006).
87. Lee, Y., Ise, T., Ha, D., Saint Fleur, A., Hahn, Y., Liu, X. F., Nagata, S., Lee, B., Bera, T. K. & Pastan, I. Evolution and expression of chimeric POTE-actin genes in the human genome. *Proc Natl Acad Sci U S A* **103**, 17885-17890 (2006).
88. Hahn, Y., Bera, T. K., Gehlhaus, K., Kirsch, I. R., Pastan, I. H. & Lee, B. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci U S A* **101**, 13257-13261 (2004).
89. Hahn, Y. & Lee, B. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics* **21 Suppl 1**, i186-194 (2005).
90. Hahn, Y. & Lee, B. Human-specific nonsense mutations identified by genome sequence comparisons. *Hum Genet* **119**, 169-178 (2006).
91. Hahn, Y., Jeong, S. & Lee, B. Inactivation of MOXD2 and S100A15A by Exon Deletion During Human Evolution. *Mol Biol Evol* (2007).
92. Rubattu, S., Stanzione, R., Di Angelantonio, E., Zanda, B., Evangelista, A., Tarasi, D., Gigante, B., Pirisi, A., Brunetti, E. & Volpe, M. Atrial natriuretic peptide gene polymorphisms and risk of ischemic stroke in humans. *Stroke* **35**, 814-818 (2004).
93. Marenholz, I., Lovering, R. C. & Heizmann, C. W. An update of the S100 nomenclature. *Biochim Biophys Acta* **1763**, 1282-1283 (2006).
94. Wolf, R., Voscopoulos, C. J., FitzGerald, P. C., Goldsmith, P., Cataisson, C., Gunsior, M., Walz, M., Ruzicka, T. & Yuspa, S. H. The mouse S100A15 ortholog parallels genomic organization, structure, gene expression, and protein-processing pattern of the human S100A7/A15 subfamily during epidermal maturation. *J Invest Dermatol* **126**, 1600-1608 (2006).
95. Webb, M., Emberley, E. D., Lizardo, M., Alowami, S., Qing, G., Alfia'ar, A., Snell-Curtis, L. J., Niu, Y., Civetta, A., Myal, Y., Shiu, R., Murphy, L. C. & Watson, P. H. Expression analysis of the mouse S100A7/psoriasin gene in skin inflammation and mammary tumorigenesis. *BMC Cancer* **5**, 17 (2005).