Chapter I: Recombination as an Evolutionary Search Strategy

Mechanisms for Evolution

The space of all proteins is very large, containing more possible sequences than there are atoms in the universe. The protein sequences we observe today were created by evolution over millions of years, yet all the proteins explored by evolution represent a very small proportion of the possible sequence space. The frequency of functional proteins among random sequences is estimated between 1 in 10^{11} (Keefe and Szostak 2001) and 1 in 10^{77} (Axe 2004). However, nature has managed to give us thousands of proteins with a plethora of different functions principally using two mechanisms to explore possible protein sequences: mutation and recombination.

The effects of mutation on proteins have been well studied, partially because mutagenesis has been used extensively to perturb protein sequence in order to study the relationship between sequence and structure or function. Most proteins are robust to random mutations and show a great deal of tolerance for single amino acid substitutions (Rennell et al. 1991; Markiewicz et al. 1994). However, as additional random substitutions are incorporated, the probability of a protein retaining function declines exponentially with the number of substitutions (Daugherty et al. 2000; Guo et al. 2004). Simulations using model proteins have suggested that the mutational tolerance of a natural protein is a property arising from evolutionary history (Taverna and Goldstein 2002). Evolved populations of model proteins tend to form “neutral networks” of proteins clustered around a sequence that is most mutationally robust or the “prototype sequence” (Bornberg-Bauer and Chan 1999). These sequences are related by single point mutations and are selectively neutral. Some structures have larger neutral networks than others, and are considered more designable. Such highly designable sequences are robust to mutation and are thought to correspond to natural protein folds (Li et al. 1996; Govindarajan and Goldstein 1996).

The effects of recombination on protein structure and function are much less well explored. Recombination has not been extensively used to perturb existing natural proteins, thus there is much less data available on the effects of recombination on protein function. Genetic studies have shown that recombination can occur within protein encoding regions (Dooner and Martinez-Ferez 1997; Feil et al. 1999; Feil et al. 2001; Fu et al. 2002). However, all of these data are drawn from genomic sequences and thus reflect extensive natural selection of the organism in addition to selective pressure at the single protein level. Studies of recombination with model proteins show nonhomologous recombination can lead to more efficient searching of sequence space than point mutation alone (Bogarad and Deem 1999; Cui et al. 2002), and that populations evolved allowing recombination are more centered around the prototype sequence (Xia and Levitt 2002).

Laboratory Evolution

Both recombination and mutation are used as tools to engineer new proteins with desirable properties. While evolution of natural proteins has taken place with large populations over millions of years, the situation in the laboratory is more limited. In laboratory evolution an improved variant must be identifiable within a population of variants, and experimental constraints limit the number of variants that can be searched in a single generation (Voigt et al. 2001a). Thus finding the improved variants within a population is critical for success. Recombination and mutation allow access to very different areas of sequence space, and depending on the desired properties one may be more beneficial than the other.

Random mutagenesis using error-prone PCR is by far the easiest method for generating diversity in directed evolution. It is inexpensive, requires no information beyond a starting protein sequence, and is usually effective at producing variants with modified properties such as increased thermostability or altered substrate specificity. However, as structural and sequence information have accumulated, site-saturation mutagenesis has become more widely used for generating diversity. Site-saturation mutagenesis seeks to bias the population of variants generated to specific areas of sequence space in order to increase the probability of identifying variants of interest. It also allows access to some mutations that cannot be searched using random mutagenesis due to the conservative nature of the genetic code.

Several different studies have shown that variant populations incorporating many mutations simultaneously are better at adaptive evolution than population of variants incorporating only a few mutations (Crameri et al. 1998; Zaccolo and Gherardi 1999; Daugherty et al. 2000). However, additional diversity generated by increasing the number of random mutations comes at a cost to the fraction of functional variants. Due to the exponential decline in functional proteins as the number of random mutations increases, the majority of sequences constructed with a high mutation rate are nonfunctional and likely not folded. Without very high throughput screening or selection techniques, identifying an improved variant in the large population is impossible.

Recombination between protein sequences makes it possible to take larger steps in sequence space without sacrificing the fraction of folded variants. Recombination of proteins sharing the same fold allows only mutations that are compatible with the backbone characteristics of the fold. Because the amino acids introduced by recombination are unlikely to have deleterious interactions with the backbone, the major contributions to whether a chimera retains function are the pairwise interactions between the amino acid substitutions introduced. Thus, mutations introduced by recombination are less likely to disturb the protein structure than random mutations (Drummond et al. 2005). Recombination of homologous proteins mimics the accumulation of neutral mutations that occur in nature, but on a shorter time scale.

Homologous recombination has been a successful strategy for creating protein variation in directed evolution studies. Proteins isolated from DNA shuffling experiments have properties that are due to the action of several mutations working synergistically (Stemmer 1994; Crameri et al. 1998). However, DNA shuffling experiments are still limited in the area of sequence space that they can explore. The variants produced are usually not significantly different than their parents, displaying more than 85% sequence identity to the closest parent (Arnold 2000). While this represents a significant step in sequence space, the diversity of natural proteins suggests that even larger jumps are possible without disrupting protein structure.

DNA shuffling is an effective strategy for recombining closely related sequences. However, as the sequence identity between the DNA sequences recombined decreases below approximately 70%, DNA shuffling becomes much less effective as the number of potential crossover sites decreases and parental genes are recovered more frequently from the reaction. In addition to limiting the sequence diversity of the proteins recombined, annealing-based methods also bias the positions of recombination points and the incorporation of different parent sequences (Joern et al. 2002). Therefore, to increase the possible sequence diversity incorporated by recombination it is necessary to find alternative recombination methods to DNA shuffling.

To overcome the limitations of annealing-based recombination methods, homology-independent techniques have been developed that can recombine sequences of any identity (Ostermeier et al. 1999a; Ostermeier et al. 1999b; Lutz et al. 2001; Sieber et al. 2001; Kawarasaki et al. 2003). However, these techniques often result in a population of chimeras that is largely unfolded. Chimeras are unfolded for two reasons in these experiments: First the methodologies incorporate a significant number of frameshifts, deletions and insertions because recombination sites are randomly generated in the DNA sequence. This results in a large proportion (>66% statistically) of nonfunctional proteins in the library. Second, even using recombination, additional diversity introduced by recombining more distantly related sequences comes with a decrease in the fraction of folded variants (Meyer et al. 2003; Ostermeier 2003); recombining more diverse sequences results in an increased number of potentially deleterious pairwise interactions (Drummond et al. 2005).

Site-Directed Recombination

Site-directed recombination seeks to circumvent many of the problems in both annealing-based and homology-independent recombination methods by choosing specific sequence blocks of the parental genes, and reassembling these blocks combinatorially to create a library of variants (Figure I-1). Genes of any sequence identity can be used, and the assembly technique incorporates few insertions, deletions and frame shifts because the recombination sites are chosen before construction and not randomly determined during the experiment (Hiraga and Arnold 2003). Additionally, site-directed recombination can be used to bias the sequence space examined to areas that are more likely to have folded sequences. Similar to the way that site-saturation mutagenesis limits the pool of possible mutants to those expected to have certain properties, site-directed recombination limits the chimeras constructed to a pool with expected properties. Structural or previous experimental data are typically used to determine sites for site-saturation mutagenesis. For site-directed recombination, a chimera's probability of retaining fold and function can be assessed *in silico* prior to construction.

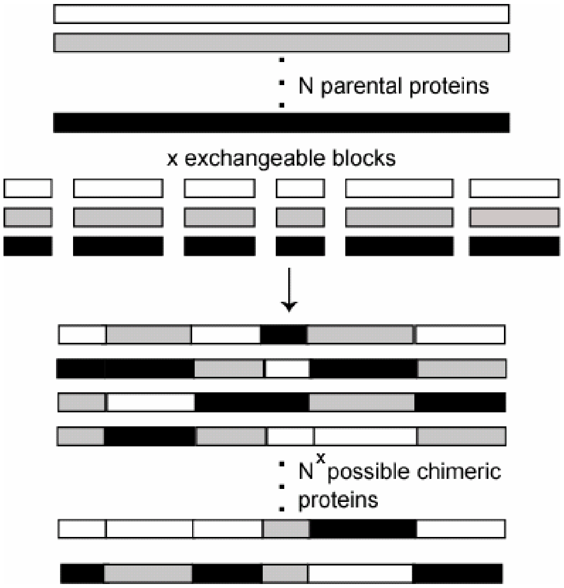

Figure I-1. Site-directed recombination recombines discrete blocks of sequence combinatorially to create a library of chimeric proteins.

Several different computational energy functions have been developed to evaluate *in silico* the probability of chimeric proteins folding (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003; Saraf et al. 2004; Hernandez and LeMaster 2005). All of these energy functions strive to predict chimera folding or functionality based on pairwise interactions. This is compatible with the understanding that pairwise side chain interactions are the largest contributor to chimera misfolding. Many energy functions identify amino acid residue pairs using three-dimensional structure information. However, the treatment of these pairs varies considerably, from simply counting clashes (Voigt et al. 2002), to mediating such clashes with biophysical information (Saraf and Maranas 2003), or mean-field calculations (Moore and Maranas 2003). Additionally, paired residues have been identified using multiple sequence alignment conservation (Saraf et al. 2004).

The scoring functions for chimeric proteins have not been tested against large, well-characterized data sets of chimeric proteins. Several have been scored against chimeras derived from directed evolution experiments (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003). However, the naïve populations from which the chimeras were selected are typically not well characterized. The lack of characterization of the naïve populations makes it difficult to determine if trends observed in functional chimeras are a result of the functional selection, or trends within the naïve population. To better understand how recombination can be used as an effective mechanism for searching sequence space, both in the laboratory and in nature, larger and better-characterized data sets are required for analysis. In this work we describe the design and creation of several large libraries of chimeric proteins. Analysis of the chimeras produced allows us to examine attributes of chimeras that contribute to folding and function, and to evaluate the tools used to predict chimeric protein folding.