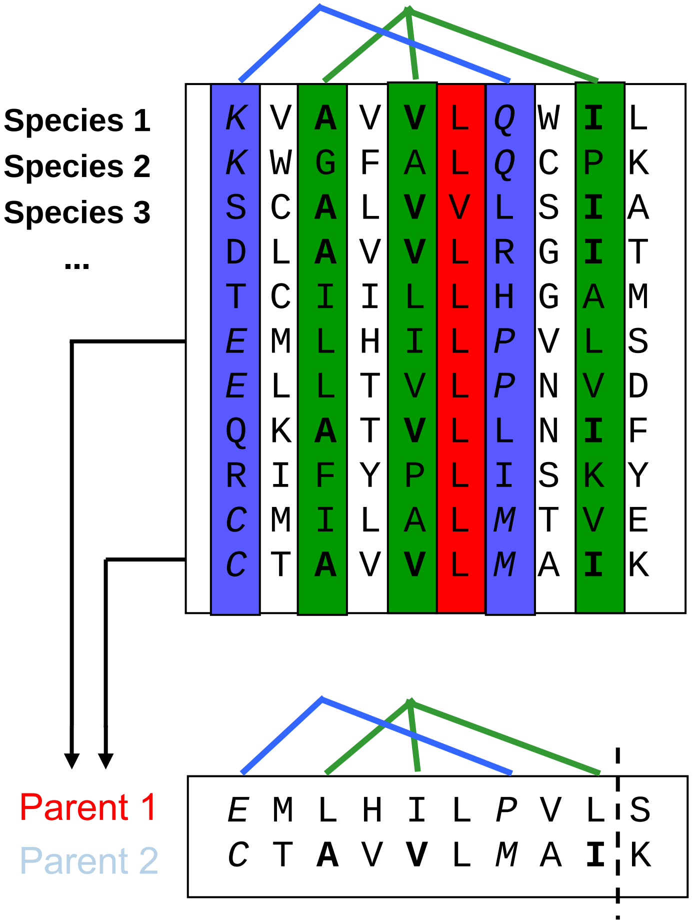
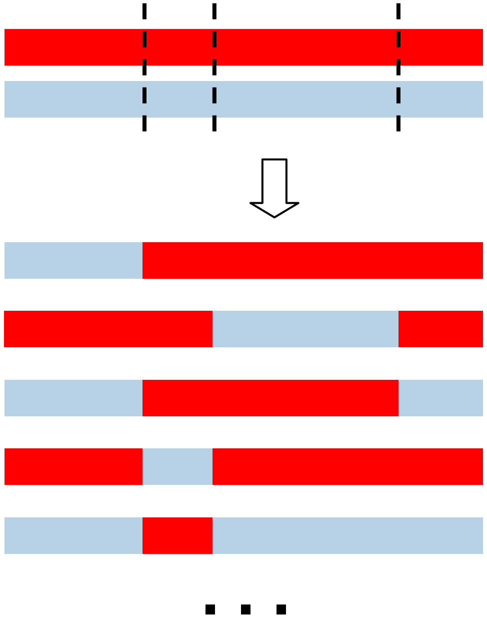
Combining new experimental & informatic tools for protein investigation & engineering



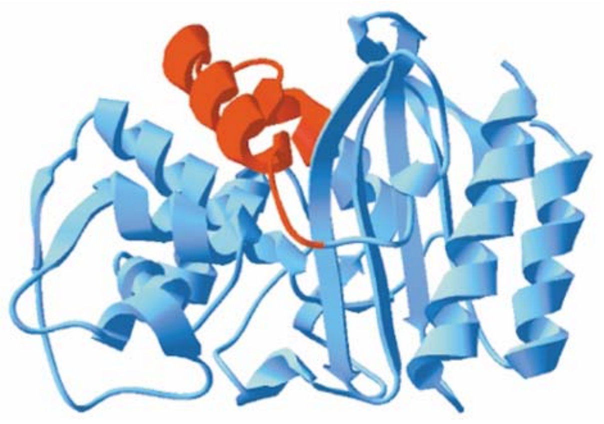
Species 1	*K*	V	**A**	V	**V**	L	*Q*	W	**I**	L
Species 2	*K*	W	G	F	A	L	*Q*	C	P	K
Species 3	S	C	**A**	L	**V**	V	L	S	**I**	A
...	D	L	**A**	V	**V**	L	R	G	**I**	T
	T	C	I	I	L	L	H	G	A	M
	E	M	L	H	I	L	*P*	V	L	S
	E	L	L	T	V	L	*P*	N	V	D
	Q	K	**A**	T	**V**	L	L	N	**I**	F
	R	I	F	Y	P	L	I	S	K	Y
	C	M	I	L	A	L	*M*	T	V	E
	C	T	**A**	V	**V**	L	*M*	A	**I**	K

Parent 1	*E*	M	L	H	I	L	*P*	V	L	S
Parent 2	*C*	T	**A**	V	**V**	L	*M*	A	**I**	K

Analyze sequence relationships & Conduct experiment planning (select parents & breakpoints)

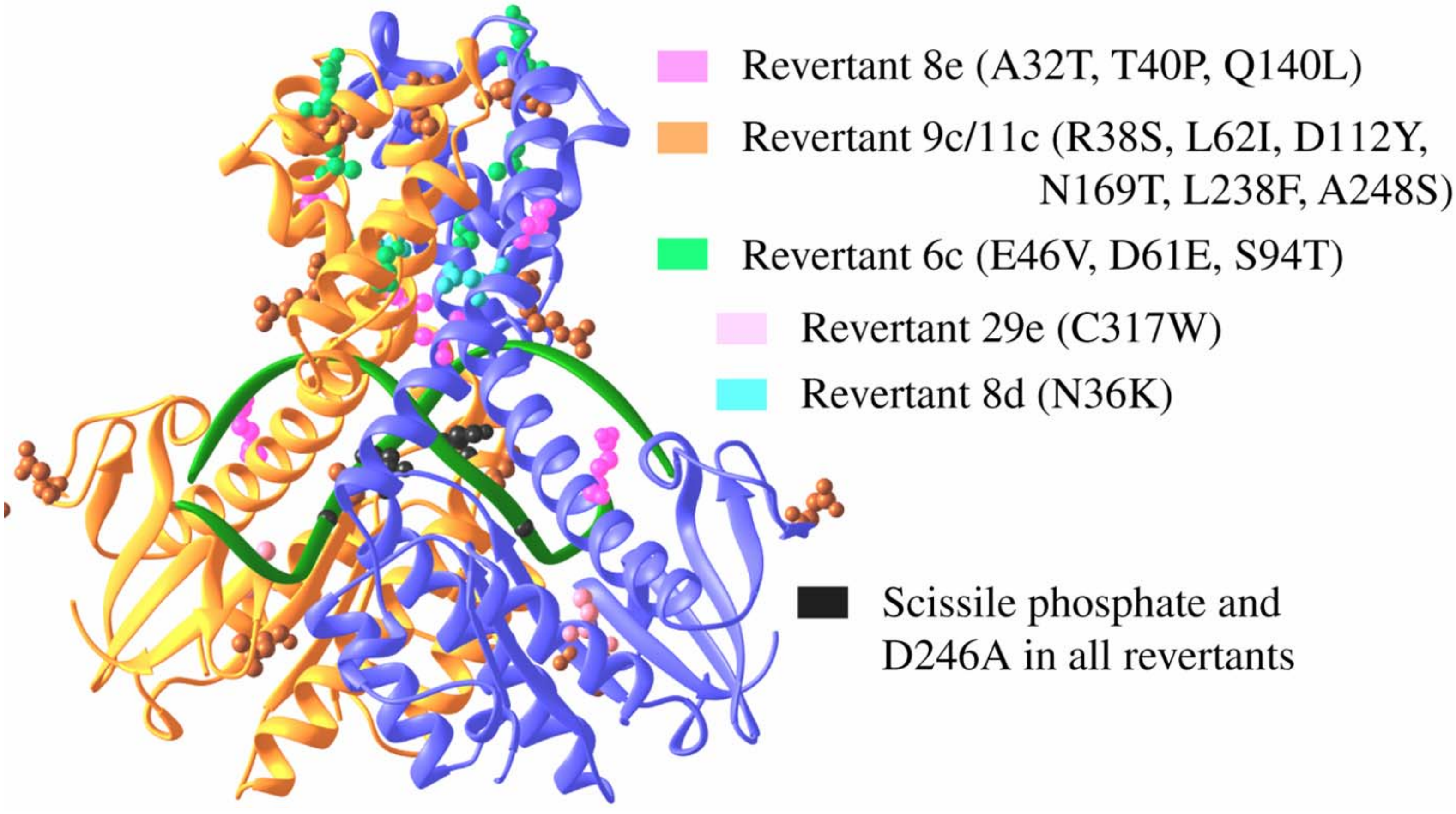


Assemble libraries of chimeric genes (SPLISO, planned DNA ligation and RoboMix, robotic mixing)



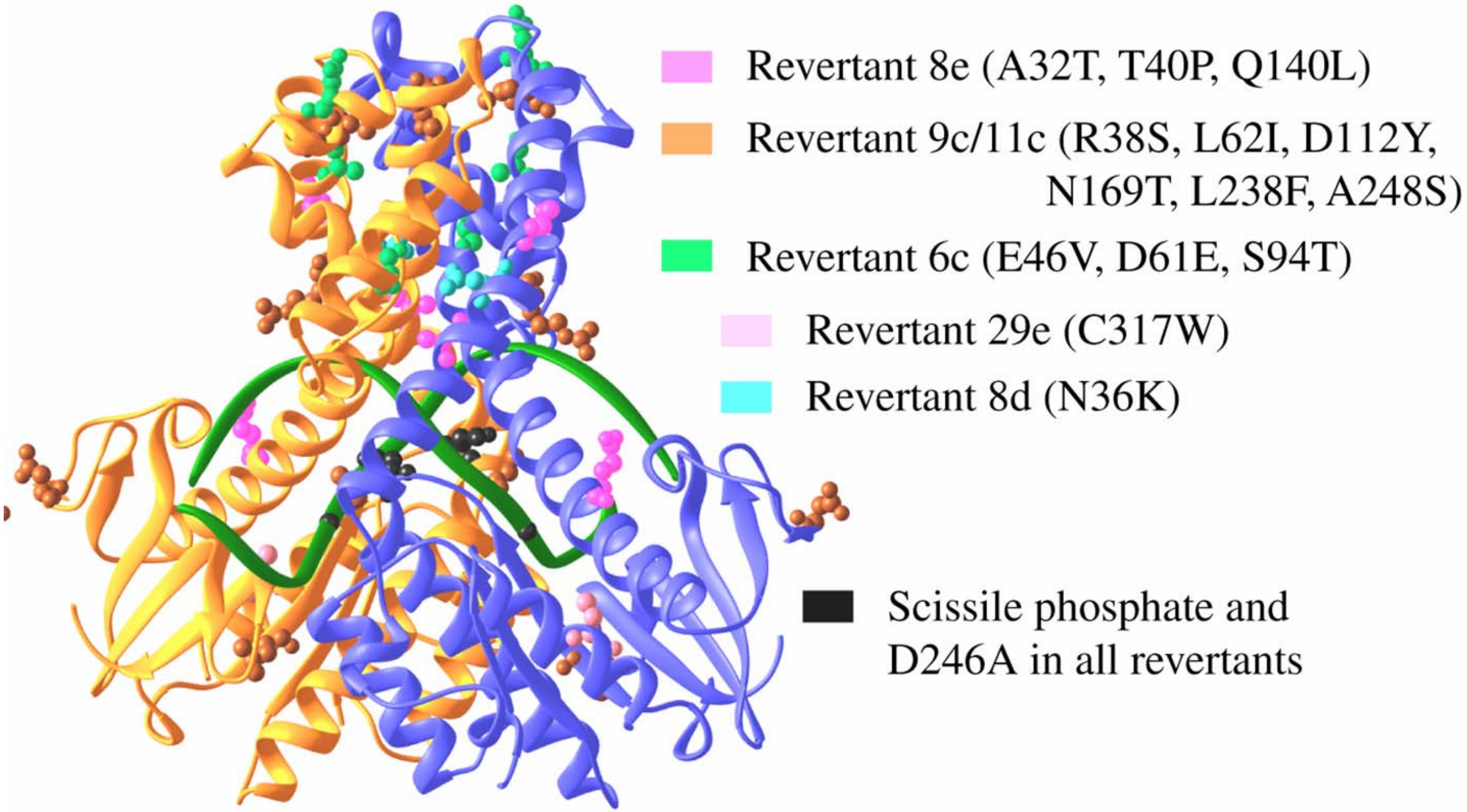
Express and analyze chimeric proteins

Motivation: Understanding the role of Structural Elements & Distant Mutants


Revertant 8e (A32T, T40P, Q140L)
Revertant 9c/11c (R38S, L62I, D112Y, N169T, L238F, A248S)
Revertant 6c (E46V, D61E, S94T)
Revertant 29e (C317W)
Revertant 8d (N36K)
Scissile phosphate and D246A in all revertants

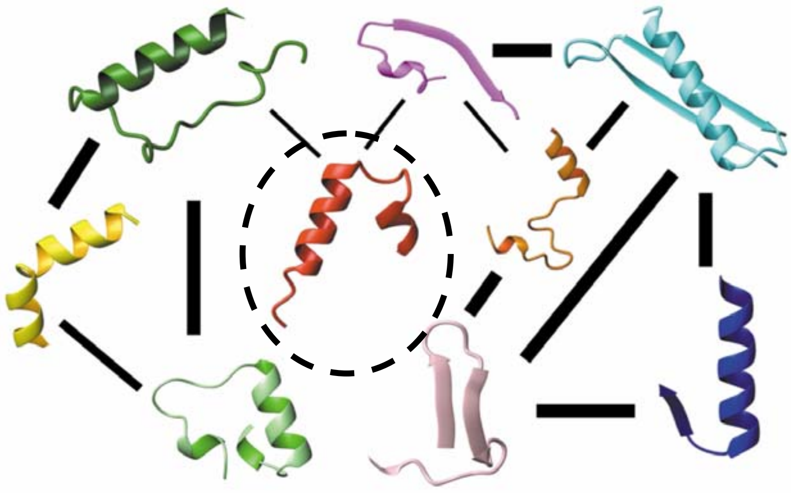
- BsoBI restriction enzyme recognizes degenerate GPuGCPyC.
- Structure suggested alterations in active site residues to ↑ specificity.
- Point mutations do ↑ specificity, but activity greatly ↓.
- Revertants selected with ↑ activity. Surprise, mutations are all over.

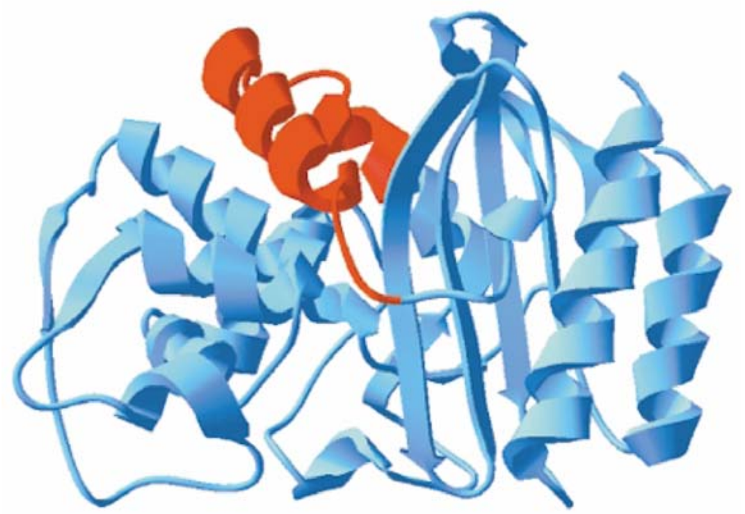
Motivation: Understanding the role of Structural Elements & Distant Mutants II


Revertant 8e (A32T, T40P, Q140L)
Revertant 9c/11c (R38S, L62I, D112Y, N169T, L238F, A248S)
Revertant 6c (E46V, D61E, S94T)
Revertant 29e (C317W)
Revertant 8d (N36K)
Scissile phosphate and D246A in all revertants

- How do distant residues have such a role?
- How do those residues (and all the others) interact to create function?
- Related to old question: Why need the entire (large) protein?
- What are the other residues and structural elements good for? What properties do they confer?

Approach: Chimera Generation, Conceptually Simple but Powerful Experiment

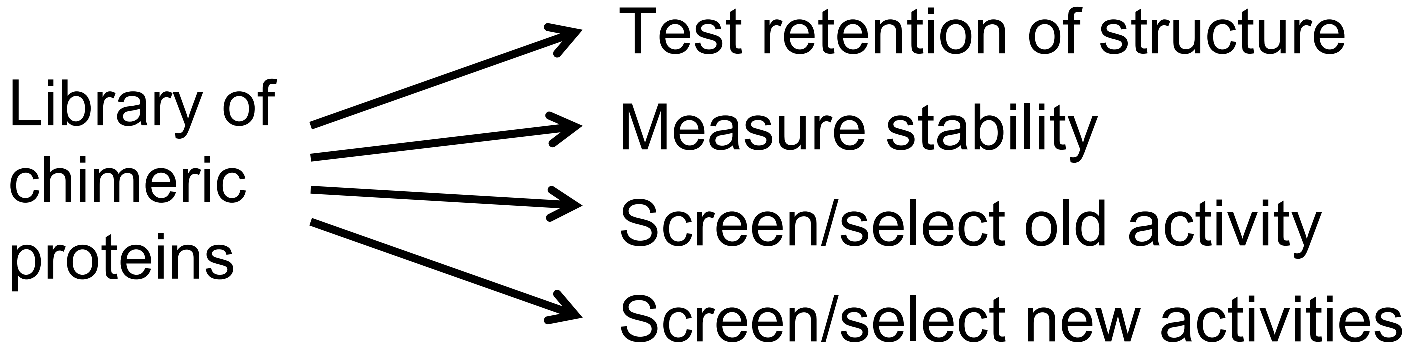




Voigt, et al., Nat. Struct. Biol. 9:553

- Figuratively, divide a protein into modules
  - Qualitatively, partially independent elements (smaller than the largely independent domains)
  - Not directly equivalent to secondary structure
- Reassemble the protein using modules from different homologs
  - Homologous pieces close enough to be roughly interchangeable
  - Better division into more independent elements -> greater interchangeability
  - Different homologous parents can provide great sequence diversity
  - Sequence diversity channeled along functional lines ("works" in one homolog)

Approach: Chimera Generation, a Flexible, Multi-Use Experiment



Library of chimeric proteins

Test retention of structure

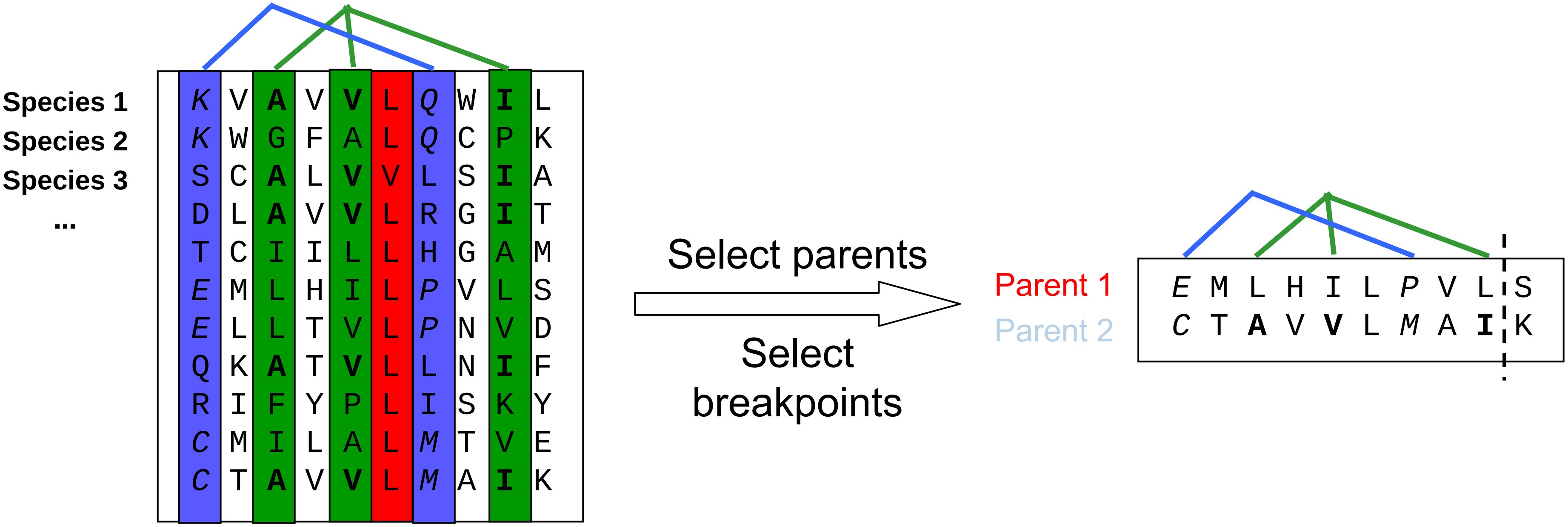
Measure stability

Screen/select old activity

Screen/select new activities

- Chimeric proteins tested to:
  - Probe the origins of structure, stability & the old “natural” activity
  - Search for new desired phenotypes (protein engineering)
- Chimera generation can be optimized to:
  - Probe structure/stability – intentional recombine between interactions suspected to lead to structure/stability *or*
  - Investigate old activity – generally maximize retention of structure and stability in library while recombining between hypothetical activity elements *or*
  - Discover new activities - compromise to combine good structure retention and stability with generating diverse new sequences

Approach: Optimization of Chimera Generation



```
K V A V V L Q W I L
K W G F A L Q C P K
S C A L V V L S I A
D L A V V L R G I T
T C I I L L H G A M
E M L H I L P V L S
E L L T V L P N V D
Q K A T V L L N I F
R I F Y P L I S K Y
C M I L A L M T V E
C T A V V L M A I K
```

```
E M L H I L P V L S
C T A V V L M A I K
```

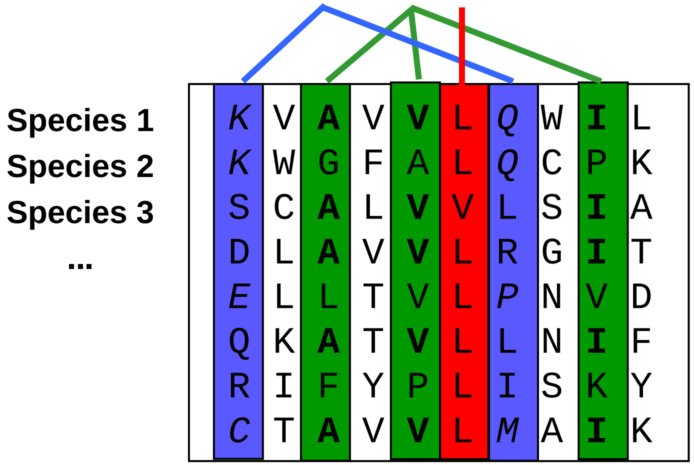
- Optimize for our goals by selecting parents and selecting recombination sites (breakpoints) -> experimental plan
  - Limits number to select/screen, #chimera = $\#\text{parents}^{\#\text{fragments}}$
  - Different parents and breakpoints yield different combinations of sequences
- How do we optimize this selection for different goals?
- And how do we actually carry out these experiments?
- New informatic and experimental technology

Optimized Chimera Generation Requires Multiple Technological Capabilities

1. Computational procedures for generating effective experimental plans (breakpoints, parents) targeted to the project goals
2. Efficient experimental procedures for assembly of chimeric genes
   a. Fragment assembly (DNA ligation)
   b. Library formation, two alternatives:
      - Mixed library
      - Each chimera in an individual vessel (well of a 96 well plate)
3. High throughput screening and selection
4. Collection and analysis of results

- I’ll talk in more detail today about 1 and 2a/b

Capability 1: Computational Analysis to Guide Chimera Planning, Overview

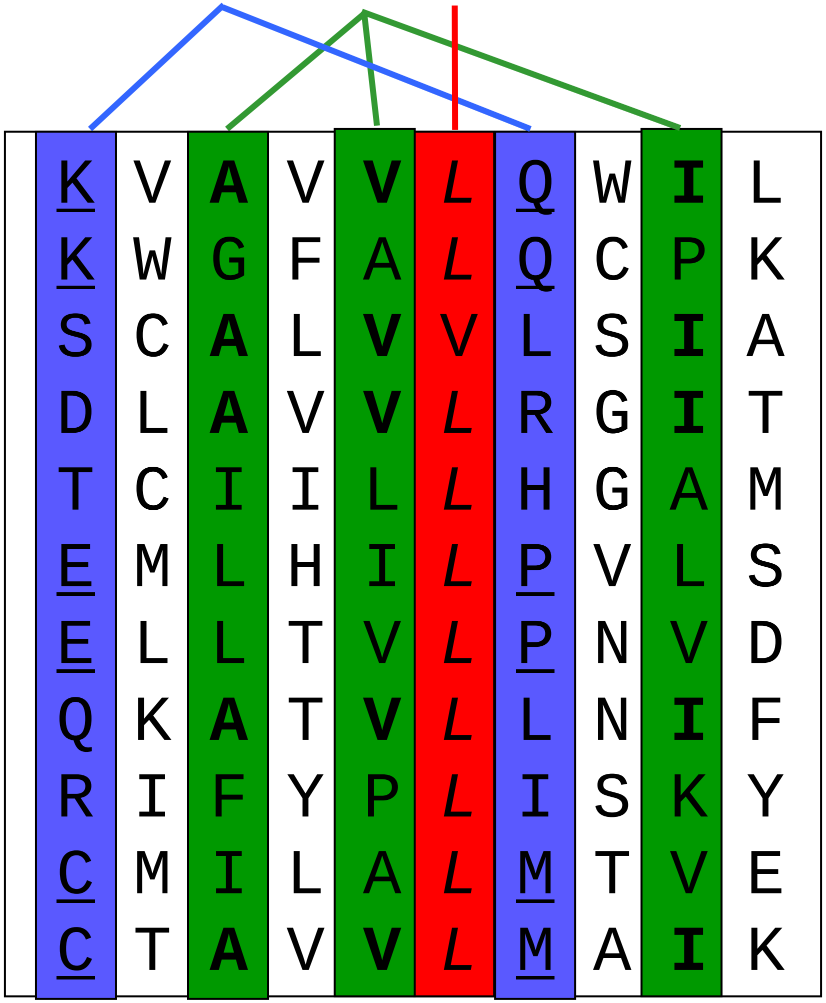


Species 1	*K*	V	**A**	V	**V**	L	*Q*	W	**I**	L
Species 2	*K*	W	G	F	A	L	*Q*	C	P	K
Species 3	S	C	**A**	L	**V**	V	L	S	**I**	A
...	D	L	**A**	V	**V**	L	R	G	**I**	T
	E	L	L	T	V	L	*P*	N	V	D
	Q	K	**A**	T	**V**	L	L	N	**I**	F
	R	I	F	Y	P	L	I	S	K	Y
	C	T	**A**	V	**V**	L	*M*	A	**I**	K

- Analysis of extant sequences in multiple sequence alignment (MSA) reveals sequence relationships (blue, red and green)
  - Relationships defined more precisely later
  - Example green dominated by Ala, Val, Ile
- For maximum structure/stability
  - Choose breakpoints to maximally preserve these groups (pairs/higher order)
    - Algorithms in Ye, et al., RECOMB 2006
  - Avoid parents that lack these groups
- For new activities
  - Choose breakpoints to compromise between preserving groups and recombining them to generate new combinations
  - Select diverse parents with both canonical and variant groups
    - Algorithms in Zheng, et al., CSB2007 & Zheng, et al., submitted RECOMB 2008

Capability 1: Analysis of multiple sequence relationships

- **Conservation** – familiar nonrandom presence of individual residues at particular positions, nearly invariant *L*
- **Correlation (coupling)** – nonrandom covariation of residues at particular positions, *KQ, EP, CM.* Mutual information between columns.
- **Hyperconservation** – nonrandom presence of groups of residues at particular sets of positions, **AVI.** Over and above conservation in individual columns.

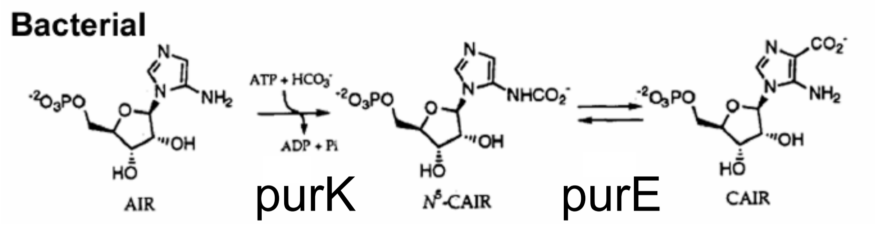


```
K V A V V L Q W I L
K W G F A L Q C P K
S C A L V V L S I A
D L A V V L R G I T
T C I I L L H G A M
E M L H I L P V L S
E L L T V L P N V D
Q K A T V L L N I F
R I F Y P L I S K Y
C M I L A L M T V E
C T A V V L M A I K
```

Ye, et al., J. Comp. Biol. 14:277

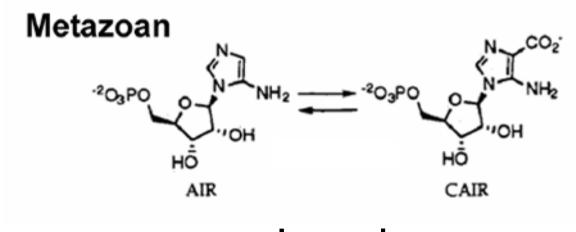
Thomas, et al., IEEE/ACM Trans. 2007

Capability 1: Multiple relationships reveal functional distinctions

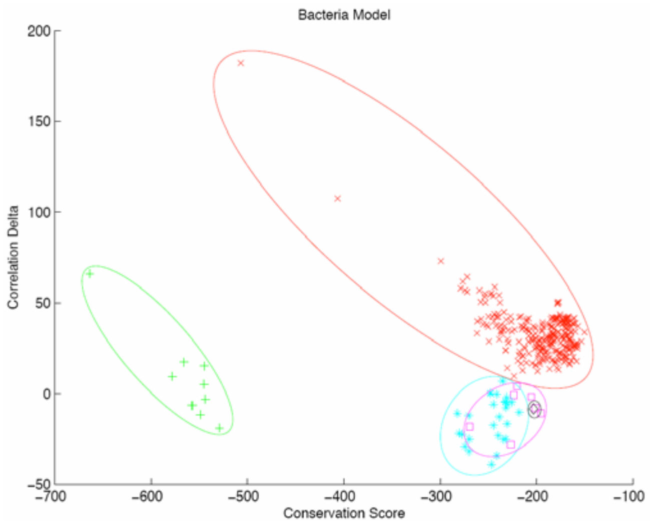


purK

purE
(mutase)



carboxylase

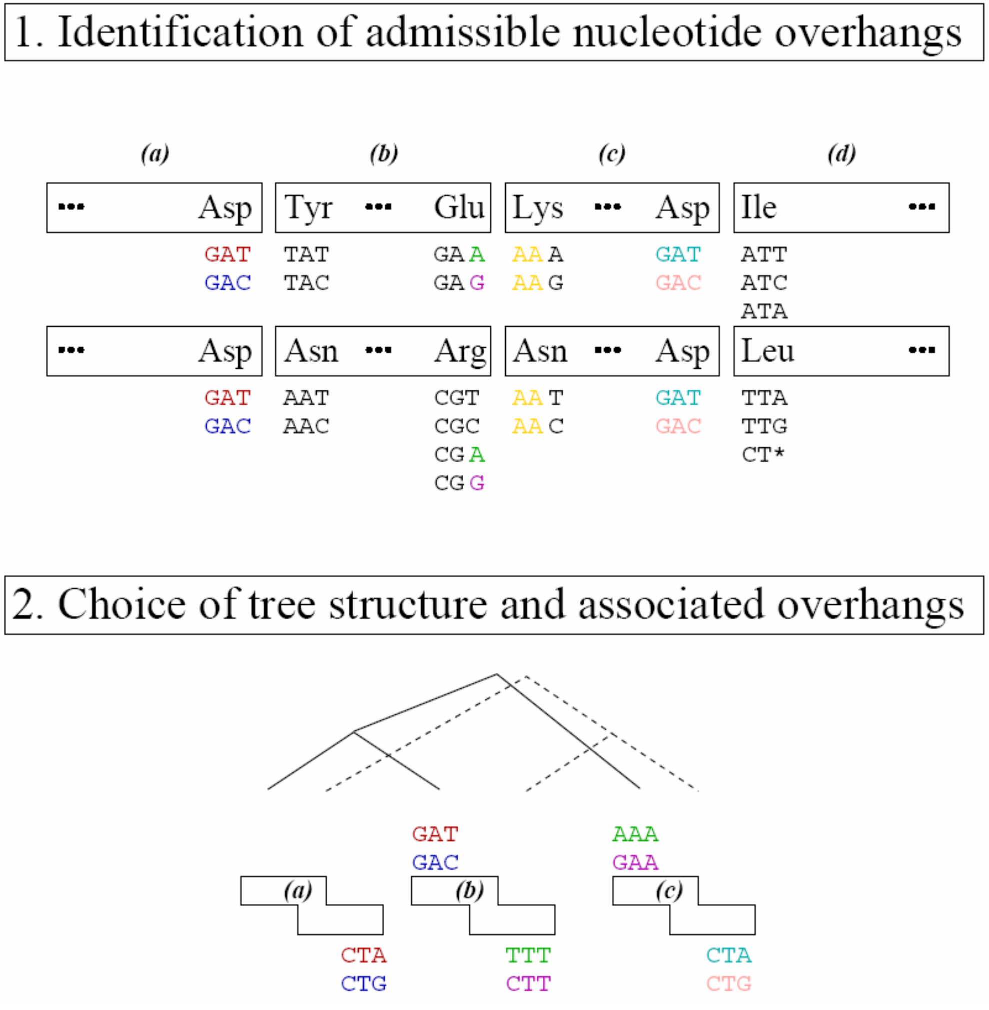


Thomas, et al., in preparation

- Project of grad student Nick Fico. Collaboration with Jo Davisson
- purE family are homologs that catalyze the same overall step in de novo purine biosynthesis
- Overall: Addition of carboxylate to AIR to form CAIR.
- Bacterial/fungal/plant have distinct mechanism from metazoans
  - Bacterial/fungal/plant mutase that move carboxylate added by purK
  - Metazoan carboxylase that adds CO_2 directly
- Demonstrates value of multiple sequence relationships. **Bacterial** and **metazoan** purE distinguished by combination of correlation and conservation, but overlap by either alone.

Capability 2a: Efficient Experimental Procedure for Chimeric Gene Assembly

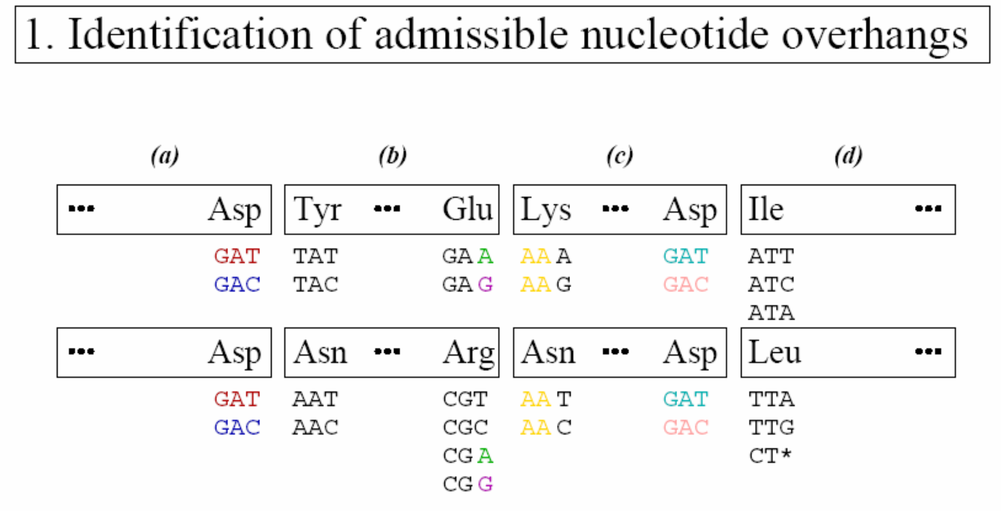
- Need mechanism for efficiently assembling fragments into chimeric genes
- SPLISO – Specific Planned LIgation of Short Overhangs
- Ligation of synthetic or PCR fragments using overhangs that do not change sequence or restrict the combinations.
- Employ type IIS restriction enzymes to generate overhangs of any sequence
- Computerized selection of best overhang sequences and assembly pathway


1. Identification of admissible nucleotide overhangs

2. Choice of tree structure and associated overhangs

Saftalov, et al., Proteins 64:629

Capability 2a: SPLISO I, identify “admissible” overhangs

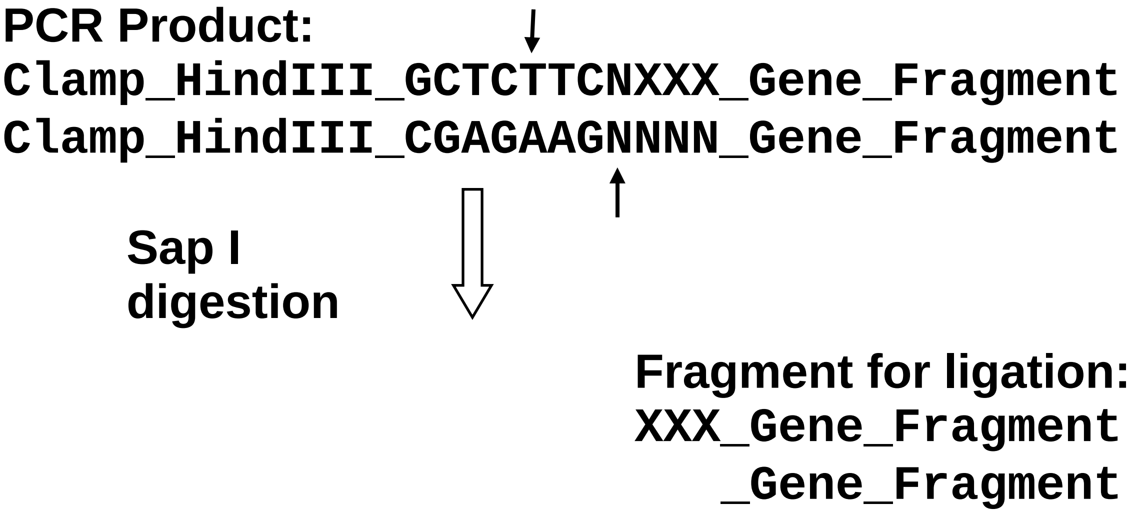


1. Identification of admissible nucleotide overhangs

(a)		(b)		(c)		(d)	
•••	Asp	Tyr •••	Glu	Lys •••	Asp	Ile	•••
	GAT GAC	TAT TAC	GAA GAG	AAA AAG	GAT GAC	ATT ATC ATA	
•••	Asp	Asn •••	Arg	Asn •••	Asp	Leu	•••
	GAT GAC	AAT AAC	CGT CGC CGA CGG	AAT AAC	GAT GAC	TTA TTG CT*	

- Use synonymous codons to identify all alternate methods of coding for the pair of amino acids spanning the breakpoint
- Construct list of overhangs that are “admissible” for all parents
  - Examples focus on 3 nt 5’ overhangs
  - Asp/Tyr & Asp/Asn easy, use either Asp codon (GAT or GAC)
  - Glu/Lys & Arg/Asn harder, use AAA or GAA, combining 3rd nt of Glu and Arg (A or G) with first two of Lys and Asn (AA)
- Computer identifies all possibilities

Practicalities: Making overhangs



PCR Product:

```
Clamp_HindIII_GCTCTTCNXXX_Gene_Fragment
Clamp_HindIII_CGAGAAGNNNN_Gene_Fragment
```

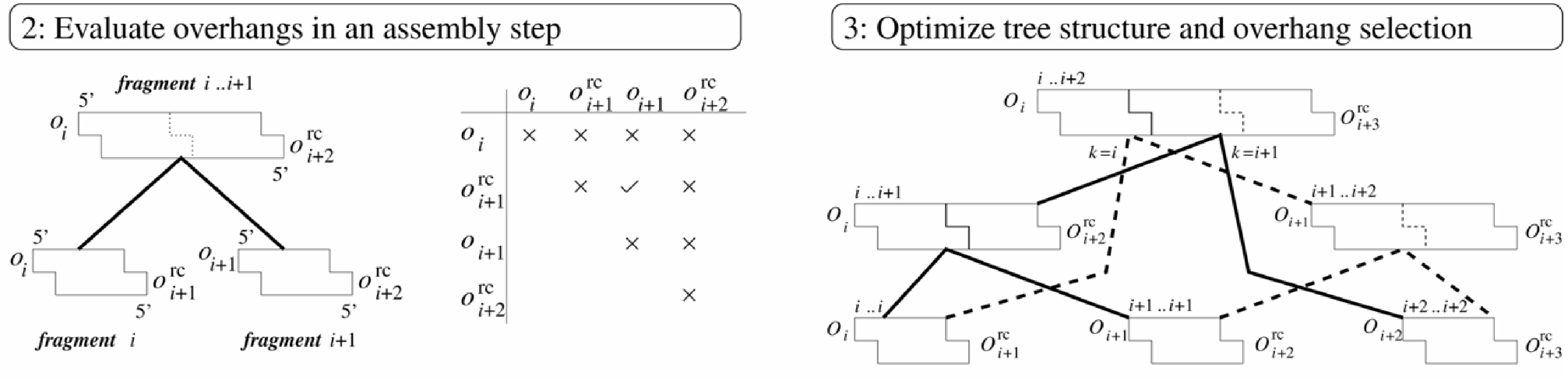
Sap I digestion

Fragment for ligation:

```
XXX_Gene_Fragment
   _Gene_Fragment
```

- Making 3 nt 5’ overhangs
  - Type IIS enzyme Sap I **GCTCTTCN**
    CGAGAAGNNNN
  - Outlined above, leaving XXX as overhang
  - Alternative enzyme sites included as backup to clone PCR fragment
- Other Type IIS enzymes leave 0-5 nt overhangs, 5’ & 3’.
- If fragment small enough to make synthetically -> any overhang
  - But longer overhangs that are admissible are more difficult to find
- Computer will output PCR primer pairs (matched T_m) or synthetic sequences for each fragment -> IDT

Capability 2a: SPLISO II, Choose best overhangs and associated assembly tree



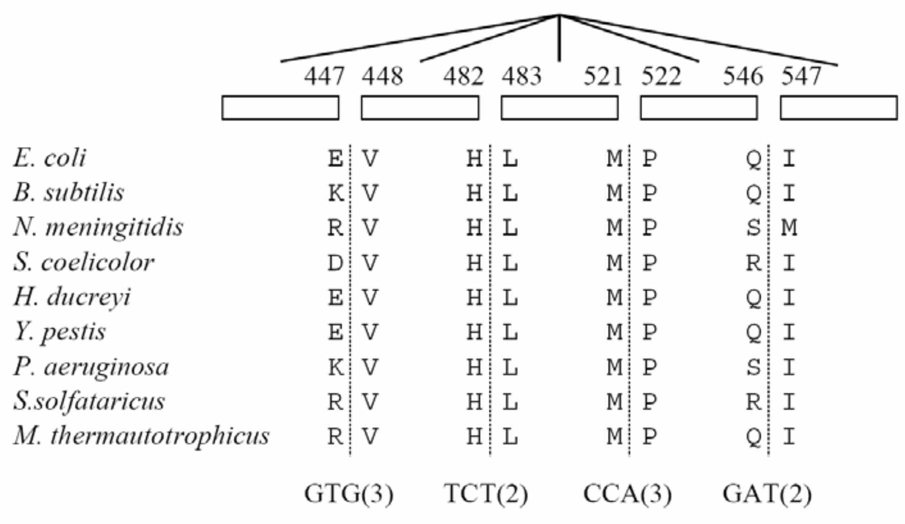
2: Evaluate overhangs in an assembly step

3: Optimize tree structure and overhang selection

	o_i	o^{rc}_{i+1}	o_{i+1}	o^{rc}_{i+2}
o_i	×	×	×	×
o^{rc}_{i+1}		×	✓	×
o_{i+1}			×	×
o^{rc}_{i+2}				×

- Score admissible overhangs in a ligation with other fragments
  - Only one ligation is desired (check mark), rest are not (x's)
  - W-C complementarity in desired ligation automatic
  - Minimize W-C complementarity in undesired pairs by varying the overhangs and/or fragments combined in each step
  - W-C complementarity above a threshold completely avoided
- Calculate best assembly pathway (tree), e.g. dotted versus solid lines
  - Different pathways combine different fragments/overhangs in each ligation
  - Earlier steps "hide" some overhangs, allows reuse of those sequences for later steps
  - Minimize tree height, # ligation steps, undesired W-C complementarity
- "Optimal substructure" allows dynamic programming to select best tree and overhangs
  - Avoids explicitly considering enormous number of possible trees multiplied by all admissible overhangs at each breakpoint

Capability 2a: SPLISO plans multi-parent, multi-breakpoint assemblies

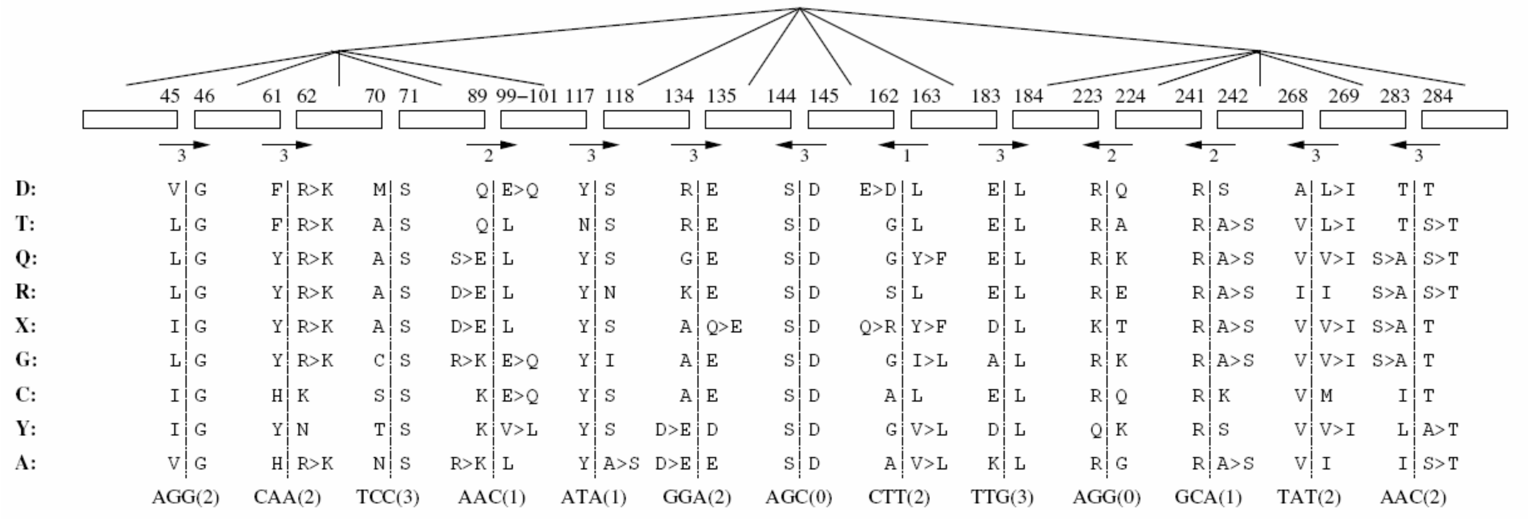


	447	448	482	483	521	522	546	547
E. coli	E	V	H	L	M	P	Q	I
B. subtilis	K	V	H	L	M	P	Q	I
N. meningitidis	R	V	H	L	M	P	S	M
S. coelicolor	D	V	H	L	M	P	R	I
H. ducreyi	E	V	H	L	M	P	Q	I
Y. pestis	E	V	H	L	M	P	Q	I
P. aeruginosa	K	V	H	L	M	P	S	I
S.solfataricus	R	V	H	L	M	P	R	I
M. thermautotrophicus	R	V	H	L	M	P	Q	I
	GTG(3)		TCT(2)		CCA(3)		GAT(2)	

One step assembly (five-way ligation) with nine PurE parents without alteration at any desired breakpoint. No overhang pair has more than 1 complementary nucleotide.

- Somewhat easy test case, sequence identity at 3 out of 4 breakpoints
- The SPLISO-determined five-way ligation not likely to go wrong
  - No more than 1 complementary nucleotide out of 3
- Testing efficiency and specificity of planned ligations in progress
- Based on well-studied principles of ligation, but planning algorithm flexible enough to allow alteration based on experimental experience
  - For example, eliminate certain overhangs that don't ligate well (possibly T rich) or restrict number of fragments ligated in one reaction

Capability 2a: Complex Assemblies with Additional Degrees of Freedom



	AGG(2)	CAA(2)	TCC(3)	AAC(1)	ATA(1)	GGA(2)	AGC(0)	CTT(2)	TTG(3)	AGG(0)	GCA(1)	TAT(2)	AAC(2)
D:	V ¦ G	F ¦ R>K	M ¦ S	Q ¦ E>Q	Y ¦ S	R ¦ E	S ¦ D	E>D ¦ L	E ¦ L	R ¦ Q	R ¦ S	A ¦ L>I	T ¦ T
T:	L ¦ G	F ¦ R>K	A ¦ S	Q ¦ L	N ¦ S	R ¦ E	S ¦ D	G ¦ L	E ¦ L	R ¦ A	R ¦ A>S	V ¦ L>I	T ¦ S>T
Q:	L ¦ G	Y ¦ R>K	A ¦ S	S>E ¦ L	Y ¦ S	G ¦ E	S ¦ D	G ¦ Y>F	E ¦ L	R ¦ K	R ¦ A>S	V ¦ V>I	S>A ¦ S>T
R:	L ¦ G	Y ¦ R>K	A ¦ S	D>E ¦ L	Y ¦ N	K ¦ E	S ¦ D	S ¦ L	E ¦ L	R ¦ E	R ¦ A>S	I ¦ I	S>A ¦ S>T
X:	I ¦ G	Y ¦ R>K	A ¦ S	D>E ¦ L	Y ¦ S	A ¦ Q>E	S ¦ D	Q>R ¦ Y>F	D ¦ L	K ¦ T	R ¦ A>S	V ¦ V>I	S>A ¦ T
G:	L ¦ G	Y ¦ R>K	C ¦ S	R>K ¦ E>Q	Y ¦ I	A ¦ E	S ¦ D	G ¦ I>L	A ¦ L	R ¦ K	R ¦ A>S	V ¦ V>I	S>A ¦ T
C:	I ¦ G	H ¦ K	S ¦ S	K ¦ E>Q	Y ¦ S	A ¦ E	S ¦ D	A ¦ L	E ¦ L	R ¦ Q	R ¦ K	V ¦ M	I ¦ T
Y:	I ¦ G	Y ¦ N	T ¦ S	K ¦ V>L	Y ¦ S	D>E ¦ D	S ¦ D	G ¦ V>L	D ¦ L	Q ¦ K	R ¦ S	V ¦ V>I	L ¦ A>T
A:	V ¦ G	H ¦ R>K	N ¦ S	R>K ¦ L	Y ¦ A>S	D>E ¦ E	S ¦ D	A ¦ V>L	K ¦ L	R ¦ G	R ¦ A>S	V ¦ I	I ¦ S>T

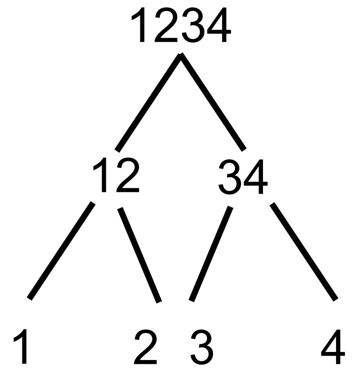
Nine parents (diverse beta-lactamases), 14 fragments using conservative substitution & breakpoint shifting allows two-height assembly

- Some experiments too complex (breakpoints too diverse, too many breakpoints, too many parents) to recombine all parents as desired
- Allow additional freedom (user-specified):
  - Conservative amino acid substitutions at the breakpoints
  - Small shifts in breakpoint location
- Additional features:
  - Select maximal possible set of parent proteins (with selection criteria).

Capability 2b: Assembly en masse or in individual wells for screening/selection

- Variants in similar experiments typically generated en masse followed by screening or selection
  - Screening requires substantial oversampling to statistically get most variants
    - Oversampling can be many-fold and assumes non-biased library
  - Selection doesn’t recover inactive variants
    - Inactive variants as informative as active ones in learning what combinations don’t work
- Alternative assembly in individual vessels allows precise recovery of all desired variants (active and inactive)
- Clint Chapple, “Actually test a hypothesis [about a single combination of fragments].”
- To do this only mix fragments from desired parents in each well
- Repeat for each desired chimera
- Robot can do this (with appropriate direction)

Capability 2b: Robotic implementation of assembly precisely mixes fragments



A simple tree

AAAA

BBBB

Two parents

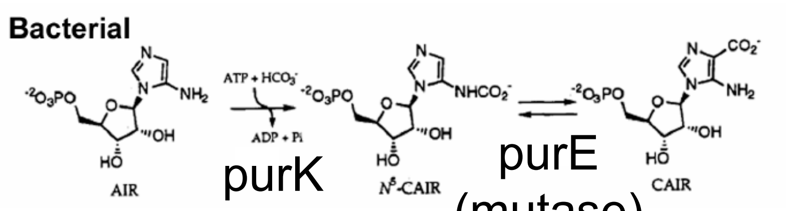
```
Well        1    2    3    4    5    6    7    8
Row A     A--- B--- -A-- -B-- --A- --B- ---A ---B
Row B     AA-- AB-- BA-- BB--
Row C     --AA --AB --BA --BB
Row D/E   AAAA AAAB AABA AABB ABAA ABAB ABBA ABBB
          BAAA BAAB BABA BABB BBAA BBAB BBBA BBBB

initially 2.0 of [1A] in well A1
initially 2.0 of [1B] in well A2
initially 2.0 of [2A] in well A3
initially 2.0 of [2B] in well A4
1.0 of well A1 into well B1
1.0 of well A3 into well B1
1.0 of well A1 into well B2
1.0 of well A4 into well B2
1.0 of well A2 into well B3
1.0 of well A3 into well B3
1.0 of well A2 into well B4
1.0 of well A4 into well B4 . . .
```

Computer generated sequence of robotic steps

Avramova, et al., J. Comb. Chem, in press

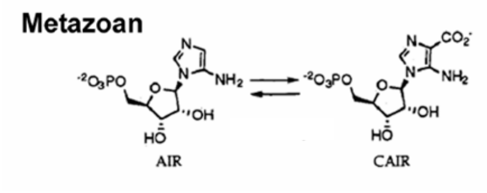
- Robots take EXCEL file aspirate & dispense.
- RoboMix generates command file from assembly tree.
- Able to generate complete set of chimera or subsets

Experimental Systems: PurE, determinants & interactions in mutase vs. carboxylase

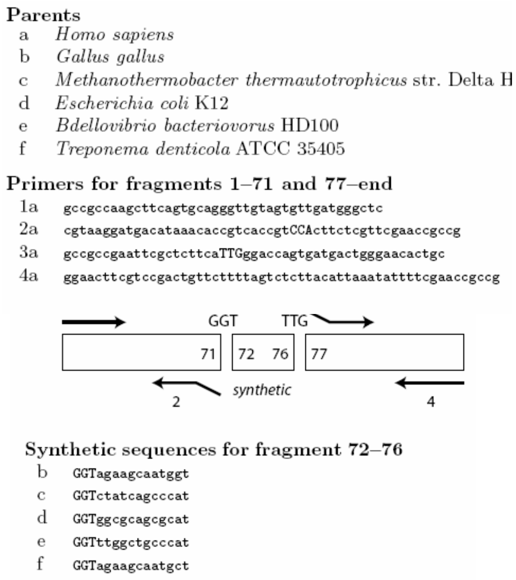


purK

purE
(mutase)



carboxylase



Parents

a	*Homo sapiens*
b	*Gallus gallus*
c	*Methanothermobacter thermautotrophicus* str. Delta H
d	*Escherichia coli* K12
e	*Bdellovibrio bacteriovorus* HD100
f	*Treponema denticola* ATCC 35405

Primers for fragments 1–71 and 77–end

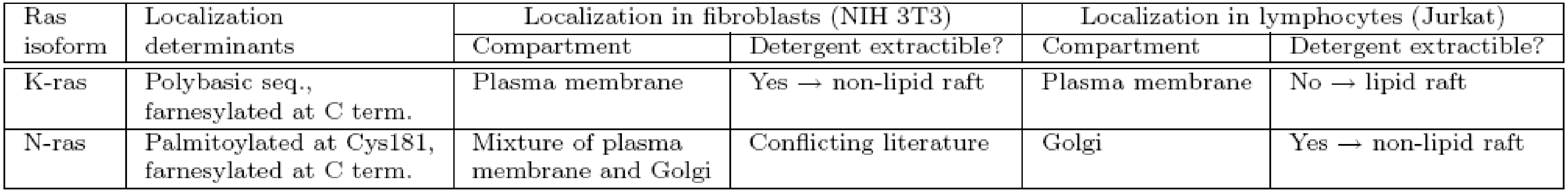
1a	gccgccaagcttcagtgcagggttgtagtgttgatgggctc
2a	cgtaaggatgacataaacaccgtcaccgtCCActtctcgttcgaaccgccg
3a	gccgccgaattcgctcttcaTTGggaccagtgatgactgggaacactgc
4a	ggaacttcgtccgactgttcttttagtctcttacattaaatattttcgaaccgccg

Synthetic sequences for fragment 72–76

b	GGTagaagcaatggt
c	GGTctatcagcccat
d	GGTggcgcagcgcat
e	GGTttggctgcccat
f	GGTagaagcaatgct

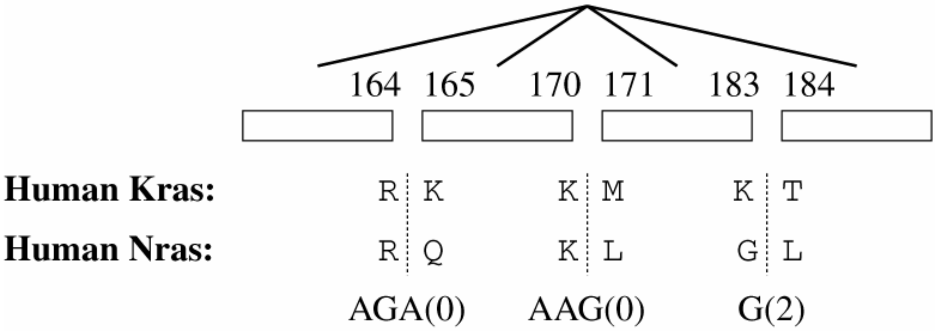
- Select carboxylase activity on purK-deficient *E. coli*
  - Modulate selection stringency by adding varying amounts of Ade
- Select mutase activity on purE$^-$, screen for no growth on purK$^-$
- Start with placing 70’s loop from several variants
  - Some evidence 70’s loop is important. Sufficient to make a mutase into carboxylase?
- Advanced: Mix fragments from several carboxylases and mutases, see which (if any) have which activity
  - Identify required individual determinants and required interactions

Experimental Systems: N-ras and K-ras, determinants & interactions determining cellular localization



Ras isoform	Localization determinants	Localization in fibroblasts (NIH 3T3)		Localization in lymphocytes (Jurkat)	
		Compartment	Detergent extractible?	Compartment	Detergent extractible?
K-ras	Polybasic seq., farnesylated at C term.	Plasma membrane	Yes → non-lipid raft	Plasma membrane	No → lipid raft
N-ras	Palmitoylated at Cys181, farnesylated at C term.	Mixture of plasma membrane and Golgi	Conflicting literature	Golgi	Yes → non-lipid raft

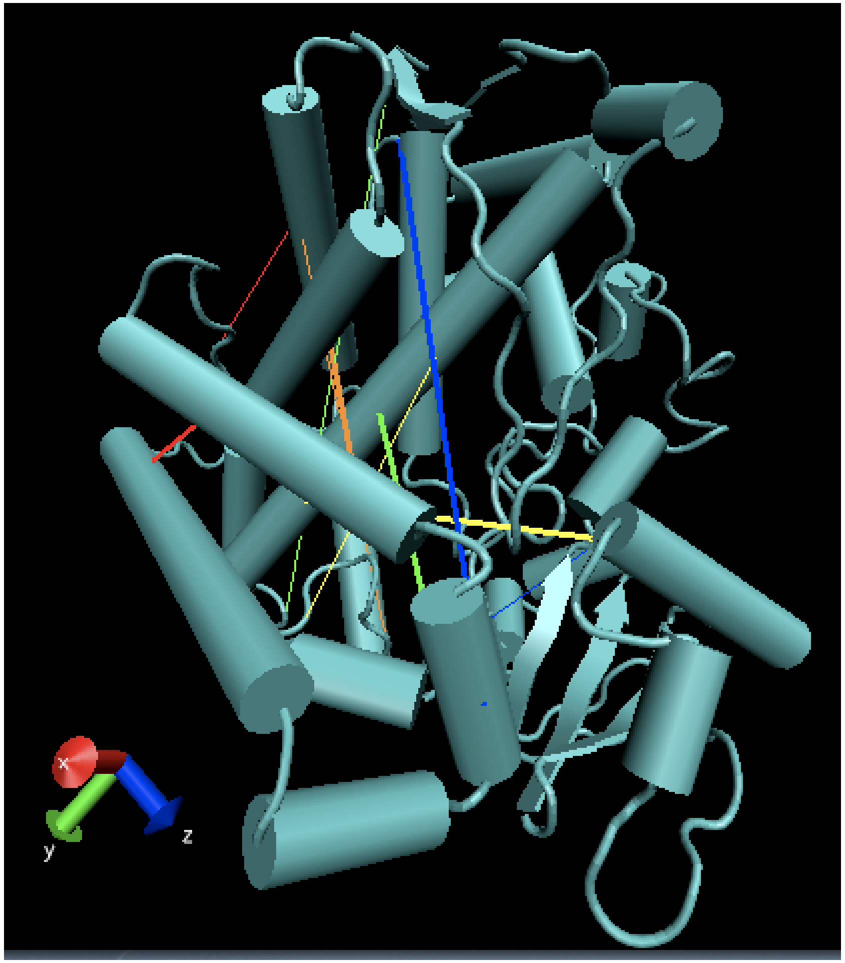
- Collaboration with Marietta Harrison and Misty Handley
- Table describes differential localization of N-/K-ras in different cell types
- What determines plasma membrane versus Golgi localization?
- Swap the C-terminal regions (and parts thereof) to identify determinants and interactions required



- Experiment plan combines 1 and 3 nt overhangs to avoid mutations at the breakpoints in divergent C-terminal region

Experimental Systems: Bioenergy, engineering of variant cytochrome P450's to modify lignin production

- Collaboration with Clint Chapple
- Lignin prevents access to sugars for fermentation into ethanol
- Develop set of lignin modification tools
- Large family of biosynthetic P450's hydroxylate lignin precursors
- Recombine them to generate novel activities
- Coupling sequence relationships from plant biosynthetic P450's determined, visualized right



Your experimental system?

?

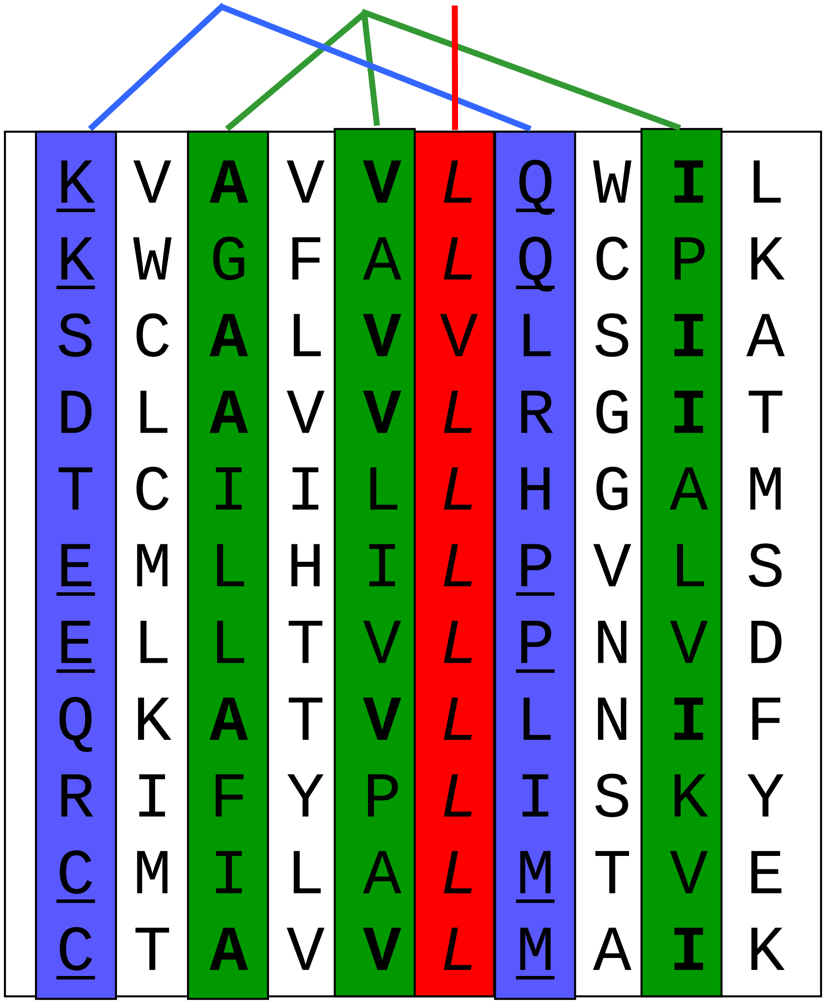
- Well-developed expression, screening and selection
- A set of homologs with which to discover sequence relationships and to serve as parents
- Do you desire a new activity?
  - Green synthetic chemistry
  - Bioenergy
  - External and internal biological modifiers
- Or have a basic investigational question?
  - Determinants **and** interactions

Bigger Picture: Epistemology of mutation *and* modeling evolution?

- Really these are old genetics questions:
  - Point and regional mutants (or swaps) -> loss of phenotype implies residue/region is a required determinant, either alone or by interaction
  - Point and regional mutants (or swaps) -> gain of phenotype implies residue/region is a “sufficient” determinant, either alone or by interaction
  - Multiple changes can tell whether determinant is acting alone or by interaction (“intramolecular epistasis”)
  - Extend to complete combinatorial->identify all interacting parts (if polymorphic)
- Are we modeling natural evolution here?
- Not sure. These are not natural alleles being recombined. Cases where similar recombination between diverse parents may be possible
  - Recombination among genes and pseudogenes
  - Recombination among viral genes in cells multiply infected with phage/viruses
  - Following promiscuous DNA transfer in prokaryotes
  - Similar situation of genes carried between species by retroviruses
- Do these reflect a significant fraction of recombination in molecular evolution?

Capability 1: Analysis of multiple sequence relationships

- **Conservation** – familiar nonrandom presence of individual residues at particular positions, nearly invariant *L*
- **Correlation (coupling)** – nonrandom covariation of residues at particular positions, *KQ, EP, CM.* Mutual information between columns.
- **Hyperconservation** – nonrandom presence of groups of residues at particular sets of positions, **AVI.** Over and above conservation in individual columns.

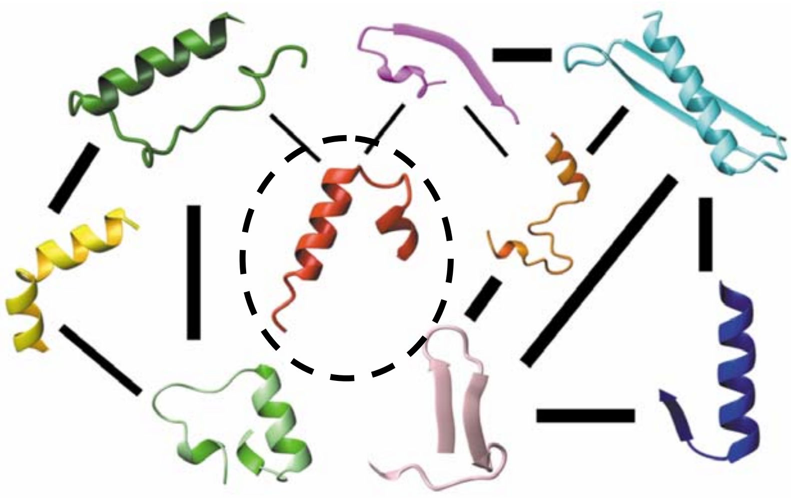


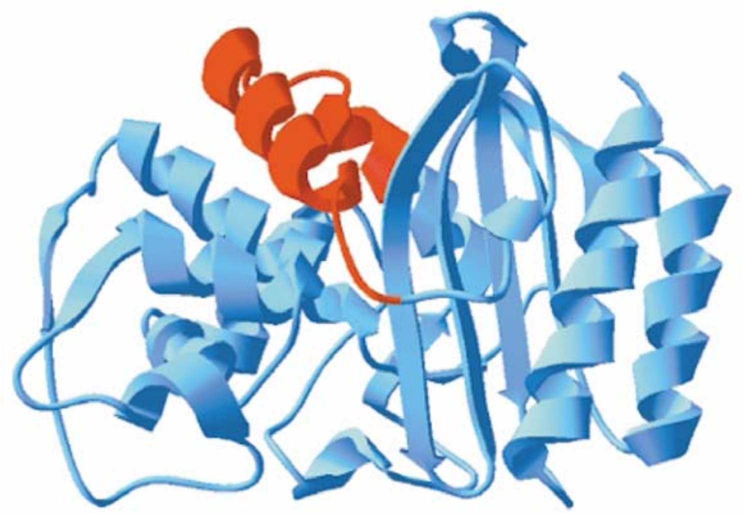
K	V	A	V	V	L	Q	W	I	L
K	W	G	F	A	L	Q	C	P	K
S	C	A	L	V	V	L	S	I	A
D	L	A	V	V	L	R	G	I	T
T	C	I	I	L	L	H	G	A	M
E	M	L	H	I	L	P	V	L	S
E	L	L	T	V	L	P	N	V	D
Q	K	A	T	V	L	L	N	I	F
R	I	F	Y	P	L	I	S	K	Y
C	M	I	L	A	L	M	T	V	E
C	T	A	V	V	L	M	A	I	K

Ye, et al., J. Comp. Biol. 14:277

Thomas, et al., IEEE/ACM Trans. 2007

Approach: Chimera Generation, Conceptually Simple but Powerful Experiment





Voigt, et al., Nat. Struct. Biol. 9:553

- Figuratively, divide a protein into modules
  - Qualitatively, partially independent elements (smaller than the largely independent domains)
  - Not directly equivalent to secondary structure
- Reassemble the protein using modules from different homologs
  - Homologous pieces close enough to be roughly interchangeable
  - Better division into more independent elements -> greater interchangeability
  - Different homologous parents can provide great sequence diversity
  - Sequence diversity channeled along functional lines ("works" in one homolog)