Evolution versus Disease:
The Calculus of Life

Olivier Lichtarge M.D., Ph.D.

Molecular and Human Genetics
Baylor College of Medicine
Houston, Texas

lichtarge@bcm.edu

A FORMAL VIEW OF DISEASE AS A PERTURBATIVE EVOLUTIONARY PROCESS

PART 1

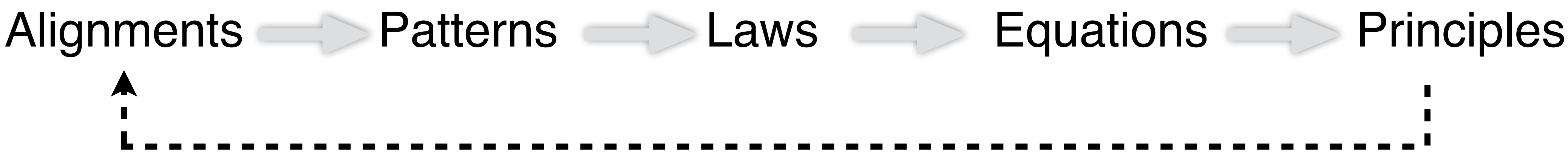
Evolutionary Trace

- *MSA*
- *Network analysis*
- *Laws of Seq-Struct-Function?*
- *Function Annotation*

PART 2

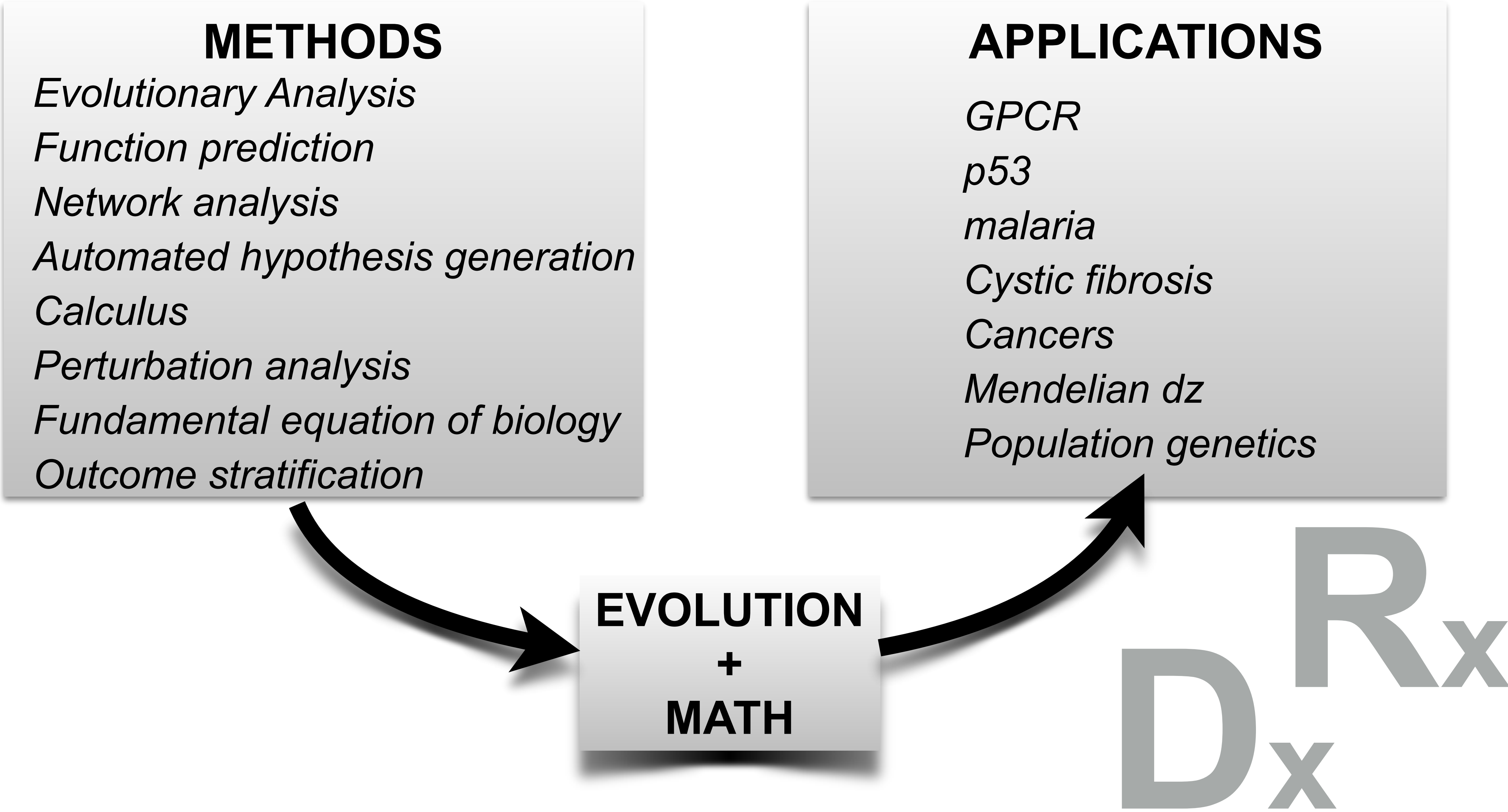
Evolutionary Action

- *Calculus*
- *Perturbation analysis*
- *Fundamental equation*
- *Outcome stratification*



Alignments → Patterns → Laws → Equations → Principles

A FORMAL VIEW OF DISEASE AS A PERTURBATIVE EVOLUTIONARY PROCESS



METHODS

- *Evolutionary Analysis*
- *Function prediction*
- *Network analysis*
- *Automated hypothesis generation*
- *Calculus*
- *Perturbation analysis*
- *Fundamental equation of biology*
- *Outcome stratification*

EVOLUTION
+
MATH

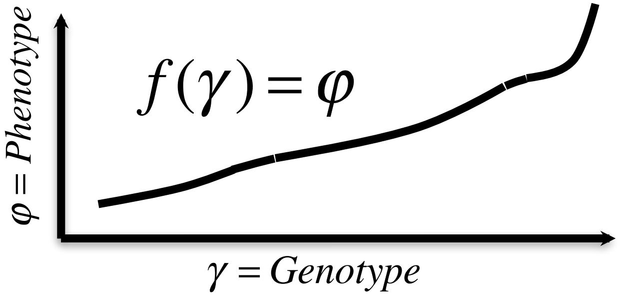
APPLICATIONS

- *GPCR*
- *p53*
- *malaria*
- *Cystic fibrosis*
- *Cancers*
- *Mendelian dz*
- *Population genetics*

FROM GENOTYPE TO PHENOTYPE

$\varphi = Phenotype$

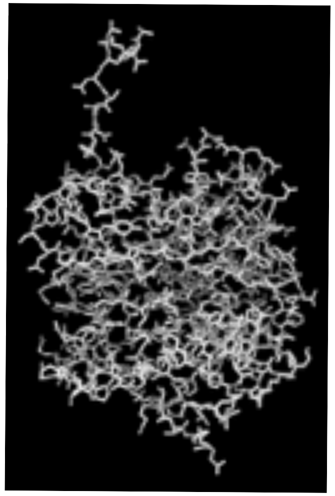
Evolutionary function



$\gamma = Genotype$

PREDICT AND DESIGN

- function
- interaction
- fitness
- mutational harm
- best therapy



COMPLEX

- Folding
- Structure
- Dynamics
- Disorder
- Modification
- Interaction
- Localization
- Degradation

MULTISCALE

- Protein
- Complexes
- Pathway
- Network
- System
- Cell
- Tissue
- Organism

GOAL 1: UNDERSTAND THE GENOTYPE-PHENOTYPE RELATIONSHIP

Lichtarge Lab

$\varphi = Phenotype$

$\gamma = Genotype$

Central Dogma

- DNA ==> RNA ==> protein

Mendel's Laws

- Segregation of parental alleles (meiosis)
- Assortment of chromosomes is random

Other Laws

- structure/function diverge exponentially
- Genes mutate inversely with expression
- Biological networks obey scale free power laws
- Evolutionary rates follow a Gaussian spread

Can we formulate biological laws in "standard form" ?

LAWS OF SCIENCE

Thermodynamics

$$\int \frac{\delta Q}{T} \geq 0$$

Inertia

$$F = d\vec{p}/dt$$

$$F_g = G\frac{m_1 m_2}{r^2}$$

$$R_{ab} - \frac{1}{2}R g_{ab} = \frac{8\pi G}{c^4}T_{ab}$$

Gravit

Schrödinger

$$\frac{p^2}{2m}|\psi\rangle + V|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle$$

Electromagnetism

$$\nabla \cdot \mathbf{D} = \rho$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$$

Fluid Dynamics

$$-\nabla p + \mu\left(\nabla^2 \mathbf{u} + \frac{1}{3}\nabla(\nabla \cdot \mathbf{u})\right) + \rho \mathbf{u} = \rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u}\right)$$

- Well-defined quantities
- Infinitesimals variations
- Conservation of some features

Is there a cornerstone to these laws?

THE POWER OF ABSTRACTION: LEAST ACTION PRINCIPLE

Thermodynamics

$$\int \frac{\delta Q}{T} \geq 0$$

Inertia

$$F = d\vec{p}/dt$$

$$F_g = G\frac{m_1 m_2}{r^2}$$

$$R_{ab} - \frac{1}{2}R g_{ab} = \frac{8\pi G}{c^4}T_{ab}$$

Gravit

Fundamental Formulas of Physics
edited by Donald H. Menzel
BASIC MATHEMATICAL FORMULAS
STATISTICS
NOMOGRAMS
PHYSICAL CONSTANTS
CLASSICAL MECHANICS
SPECIAL THEORY OF RELATIVITY
THE GENERAL THEORY OF RELATIVITY
HYDRODYNAMICS AND AERODYNAMICS
BOUNDARY VALUE PROBLEMS IN MATHEMATICAL PHYSICS
HEAT AND THERMODYNAMICS
STATISTICAL MECHANICS
KINETIC THEORY OF GASES
VISCOSITY, THERMAL CONDUCTION, AND DIFFUSION
ELECTROMAGNETIC THEORY
ELECTRONICS
SOUND AND ACOUSTICS
VOLUME ONE

Schrödinger

$$\frac{\vec{p}^2}{2m}|\psi\rangle + V|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle$$

Electromagnetism

$$\nabla\cdot\mathbf{D} = \rho$$

$$\nabla\cdot\mathbf{B} = 0$$

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$$

$$\nabla\times\mathbf{H} = \mathbf{J} + \frac{\partial\mathbf{D}}{\partial t}$$

Fluid Dynamics

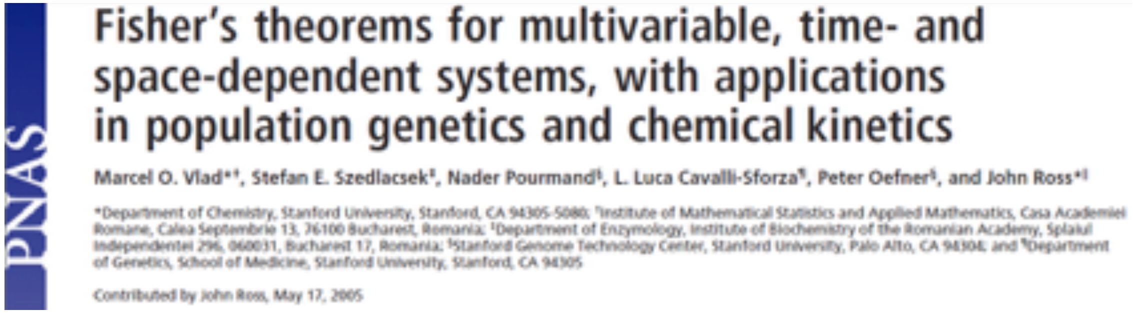
$$-\nabla p + \mu\left(\nabla^2\mathbf{u} + \frac{1}{3}\nabla(\nabla\cdot\mathbf{u})\right) + \rho\mathbf{u} = \rho\left(\frac{\partial\mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right)$$

$$\partial\int_{t_1}^{t_2} L(q,\dot{q},t)\,dt = 0$$

- A variational principle
- Yields the equations of motion
- **Noether's Theorem**: Differentiable symmetries of action are equivalent to conservation law

Perhaps we can identify and use biological variational principles ?

BIOLOGICAL VARIATIONAL PRINCIPLES — OLD QUESTIONS



PNAS

Fisher's theorems for multivariable, time- and space-dependent systems, with applications in population genetics and chemical kinetics

Marcel O. Vlad[*,†], Stefan E. Szedlacsek[‡], Nader Pourmand[§], L. Luca Cavalli-Sforza[¶], Peter Oefner[§], and John Ross[*,‖]

[*]Department of Chemistry, Stanford University, Stanford, CA 94305-5080; [†]Institute of Mathematical Statistics and Applied Mathematics, Casa Academiei Romane, Calea Septembrie 13, 76100 Bucharest, Romania; [‡]Department of Enzymology, Institute of Biochemistry of the Romanian Academy, Splaiul Independentei 296, 060031, Bucharest 17, Romania; [§]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304; and [¶]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305

Contributed by John Ross, May 17, 2005

"…there is an analogy between Fisher's theorem and the second law of thermodynamics.
…it is not clear whether Fisher's theorem is connected in some way to an extremum entropy principle."

Research program

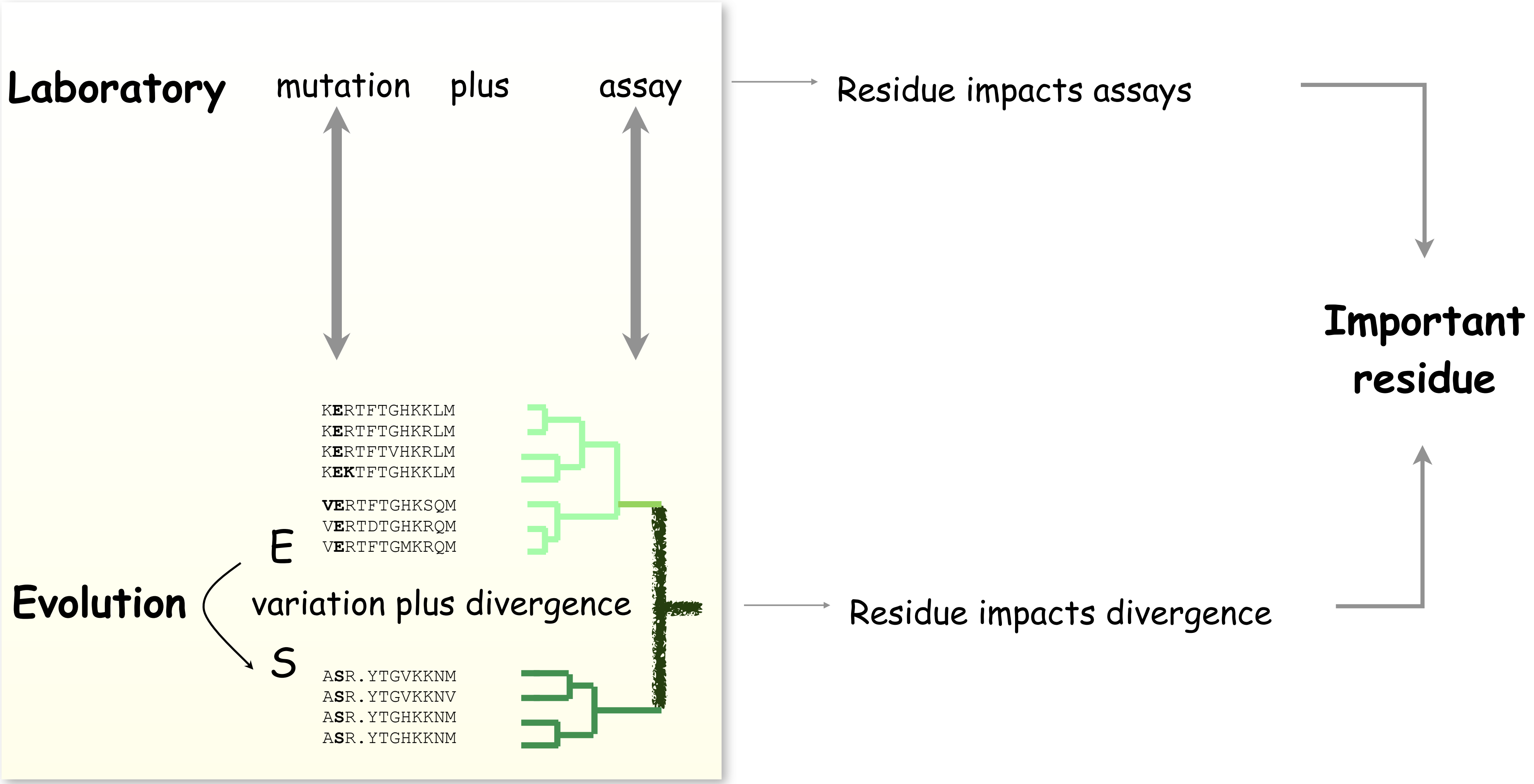
FIND
- well-defined quantities
- linked by infinitesimal variations
- and a variational principle

PREDICT/TEST
- molecular function — population genetics
- morbidity — mortality — therapy of human diseases

Part 1

Evolutionary Trace (ET)
and
Determinants of Protein Function

HOW DO PROTEINS WORK?



Laboratory mutation plus assay

→ Residue impacts assays

K**E**RTFTGHKKLM
K**E**RTFTGHKRLM
K**E**RTFTVHKRLM
K**EK**TFTGHKKLM

VERTFTGHKSQM
V**E**RTDTGHKRQM
V**E**RTFTGMKRQM

A**S**R.YTGVKKNM
A**S**R.YTGVKKNV
A**S**R.YTGHKKNM
A**S**R.YTGHKKNM

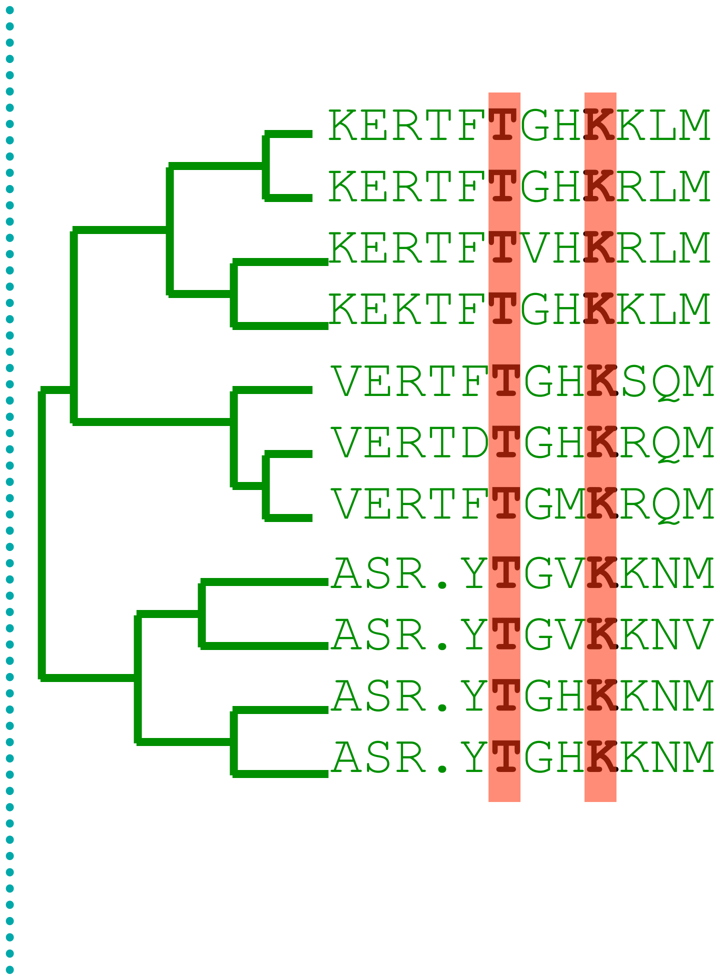
Evolution E / S variation plus divergence

→ Residue impacts divergence

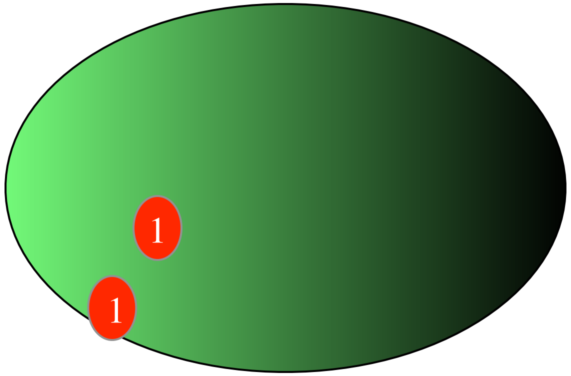
Important residue

RELEVANT PATTERNS OF VARIATIONS

1 Branch *Rank 1*



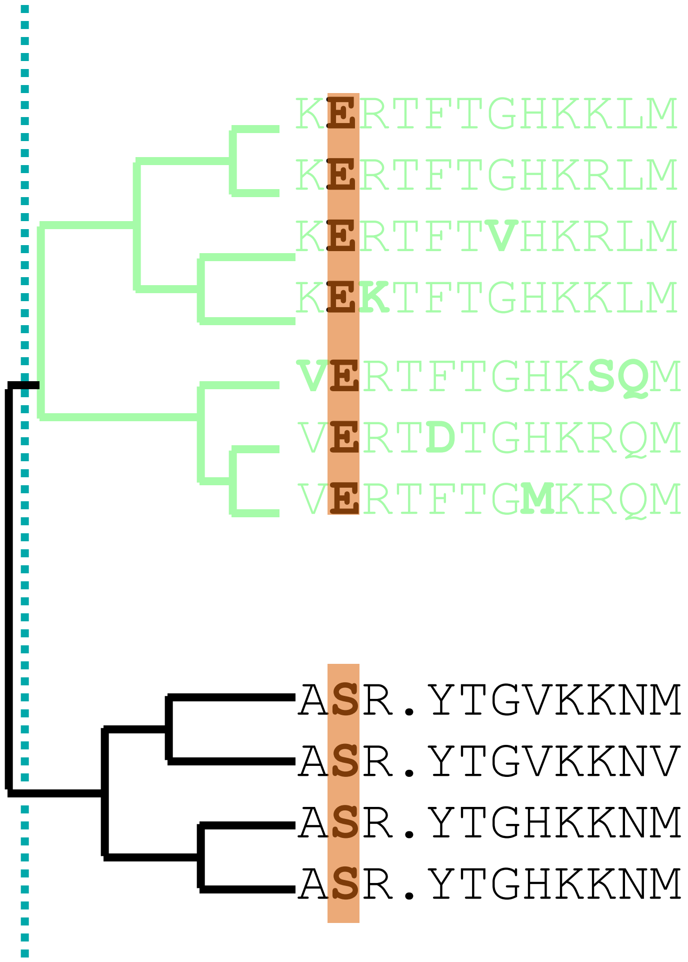
KERTFTGHKKLM
KERTFTGHKRLM
KERTFTVHKRLM
KEKTFTGHKKLM
VERTFTGHKSQM
VERTDTGHKRQM
VERTFTGMKRQM
ASR.YTGVKKNM
ASR.YTGVKKNV
ASR.YTGHKKNM
ASR.YTGHKKNM

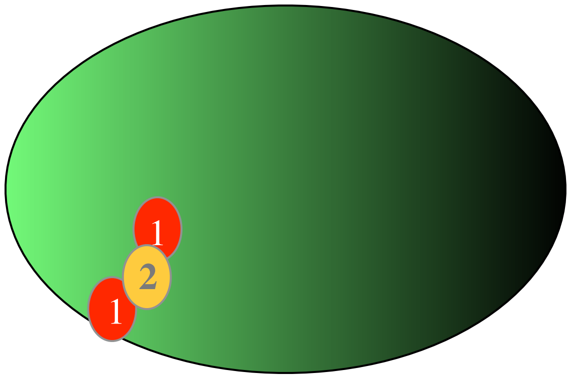


Variation pattern consistent with 1 branch

RELEVANT PATTERNS OF VARIATIONS

2 Branches *Rank 2*

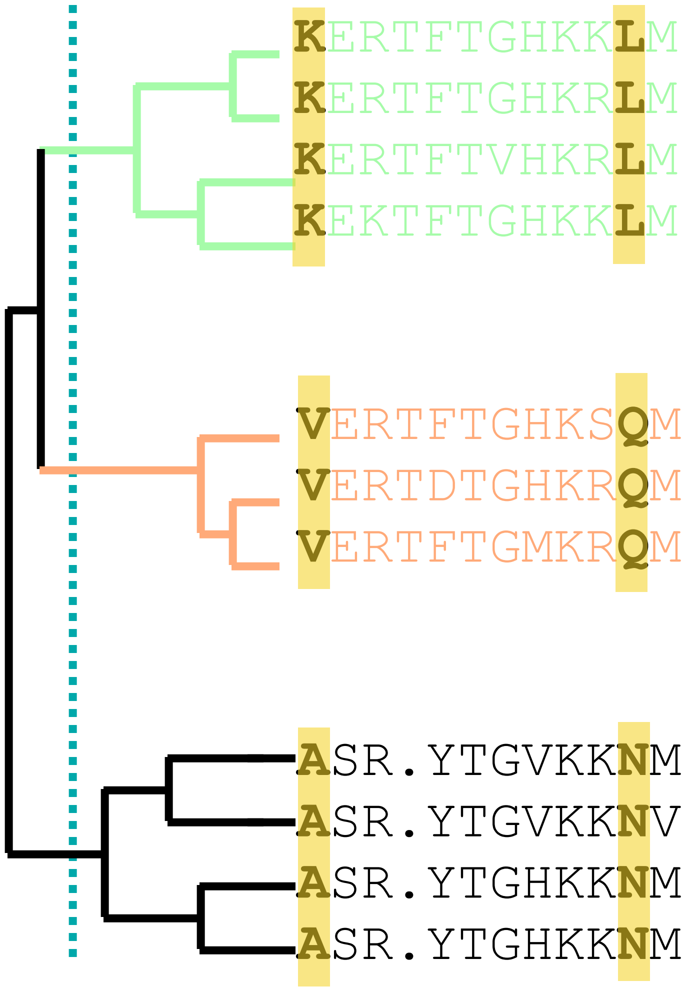


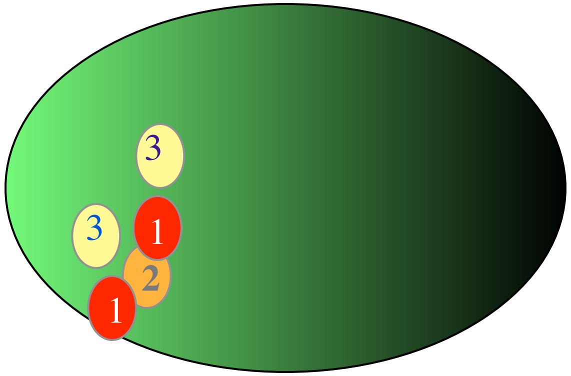


Pattern consistent with 2 branches

RELEVANT PATTERNS OF VARIATIONS

3 Branches *Rank 3*

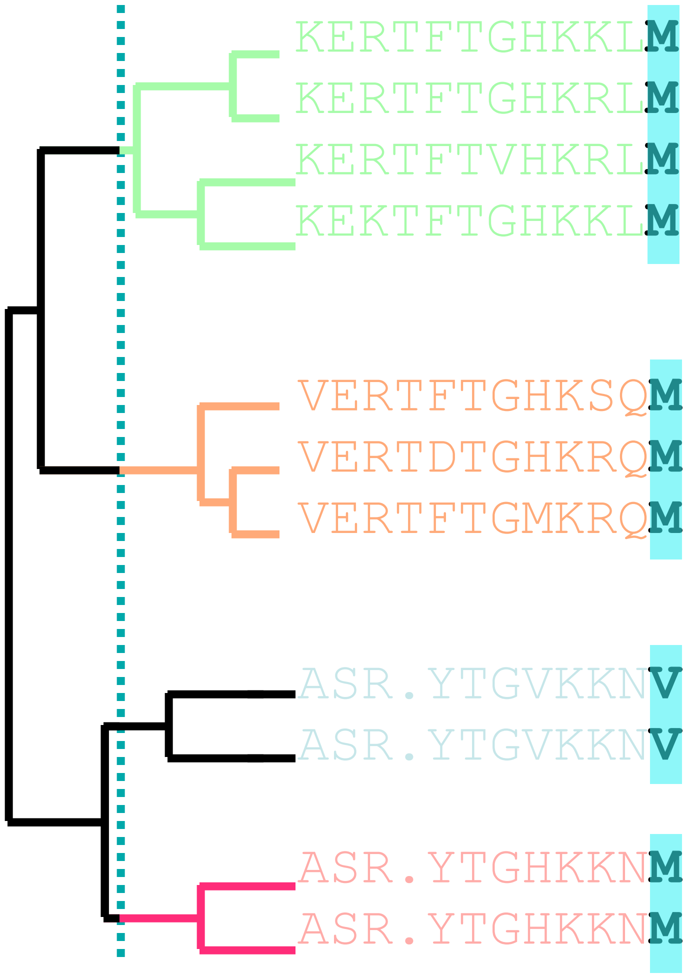


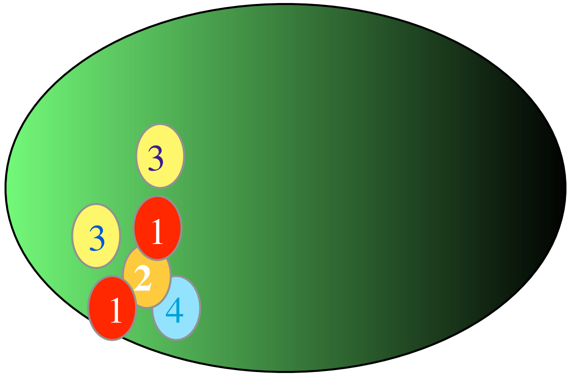


Pattern consistent with node 3 branches

EVOLUTIONARY TRACE 1.0

4 Branches *Rank 4*



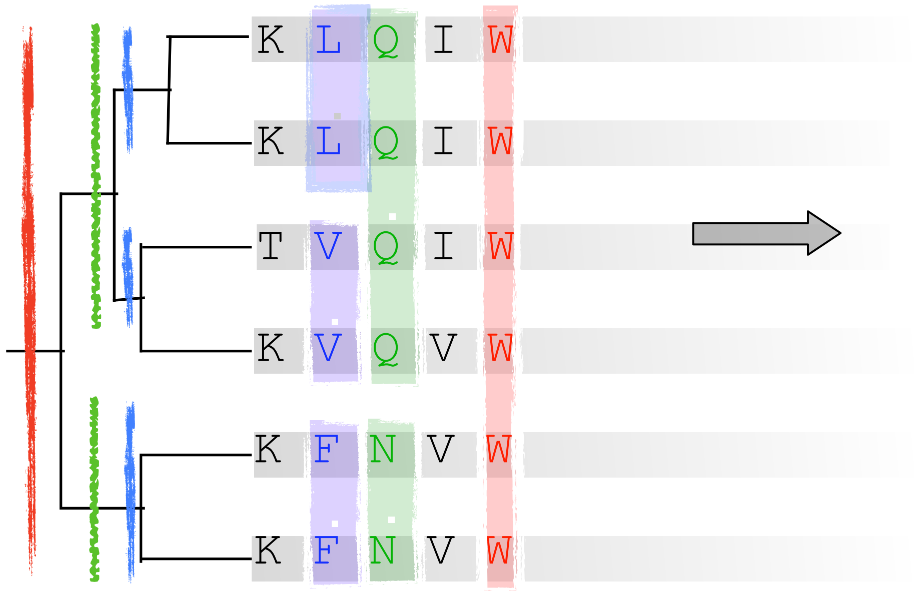


Rank of a residue:
the first node at which it varies no further

Top-ranked residues' variations are
- fixed earlier in evolution
- linked to greater evolutionary differences

Lichtarge *J. Mol. Biol.* 1996, 1997

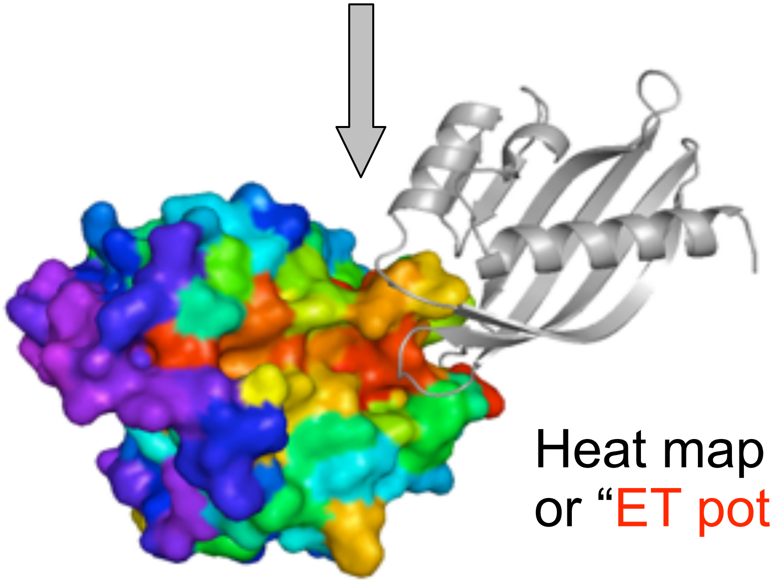
EVOLUTIONARY TRACE 2.0



K L Q I W
K L Q I W
T V Q I W
K V Q V W
K F N V W
K F N V W

*rv*ET

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \{ - \sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \}$$



Heat map of ET rank
or “ET potential”

Match evolutionary with sequence *variations*

Big *divergences* suggest important positions

Lichtarge *J. Mol. Biol.* 1996, 1997

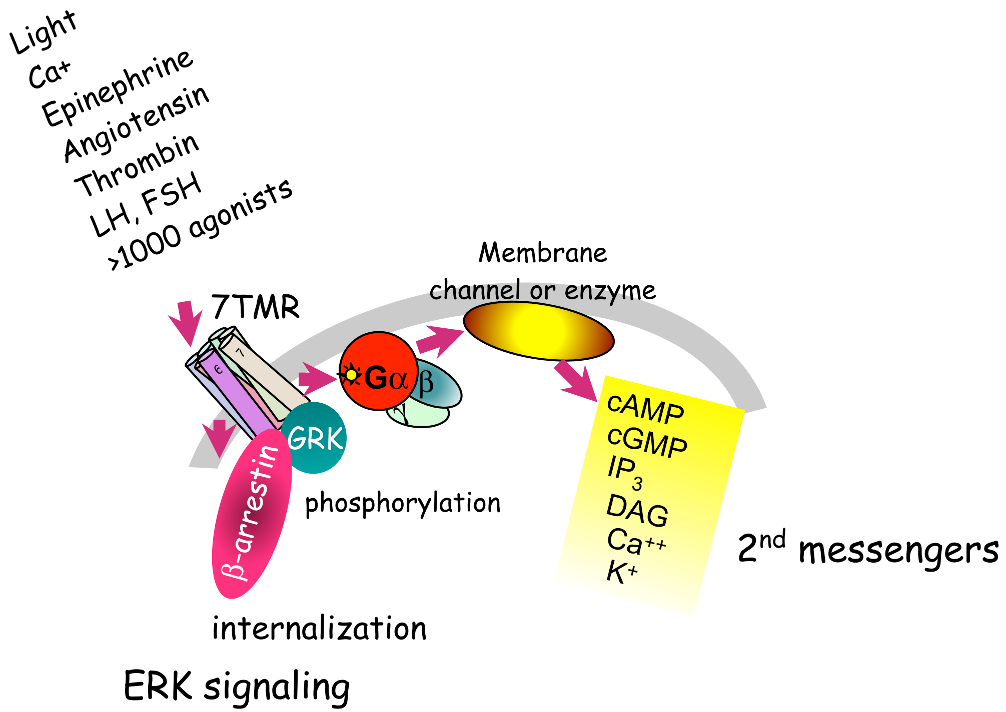
Madabushi *J. Mol. Biol.* 2002
Yao *J. Mol. Biol.* 2003
Mihalek *J. Mol. Biol.* 2004

ENGINEERING SIGNALING SPECIFICITY

G Protein-Coupled Receptors Example

RATIONAL RE-DESIGN OF GPCR SIGNALING

Make a dopamine receptor respond to serotonin


Light
Ca+
Epinephrine
Angiotensin
Thrombin
LH, FSH
>1000 agonists

7TMR

Membrane
channel or enzyme

phosphorylation

internalization

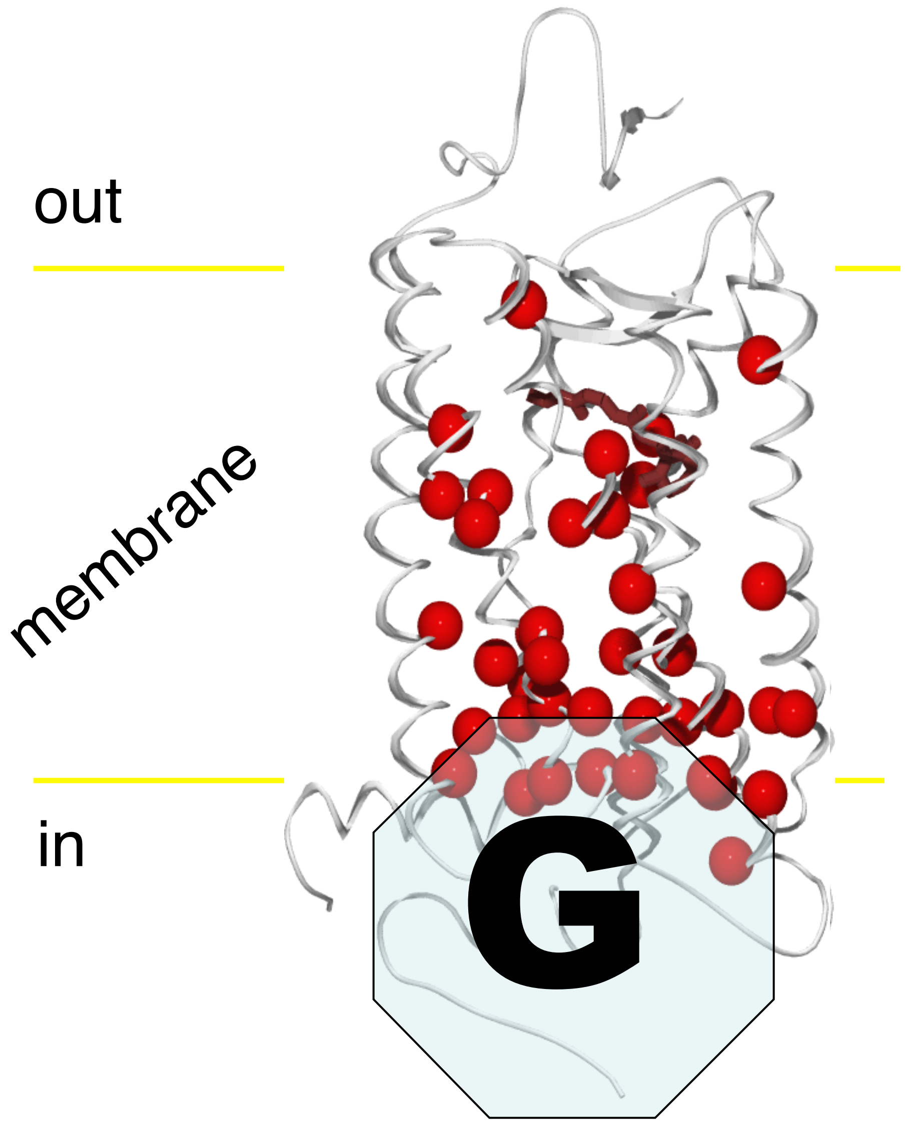
ERK signaling

2nd messengers

CNS diseases of dopamine
ADHD
Tourette’s
Parkinson’s
Social phobia
Schizophrenia
Neuroleptic malignant syndrome
Recreational drug and alcohol use

CNS diseases of serotonin
Anxiety/Mood
Learning/Memory/Perception
Sleep/Thermoregulation//Appetite
Cognition/Behavior

GPCR DETERMINANTS



Common switch mechanism

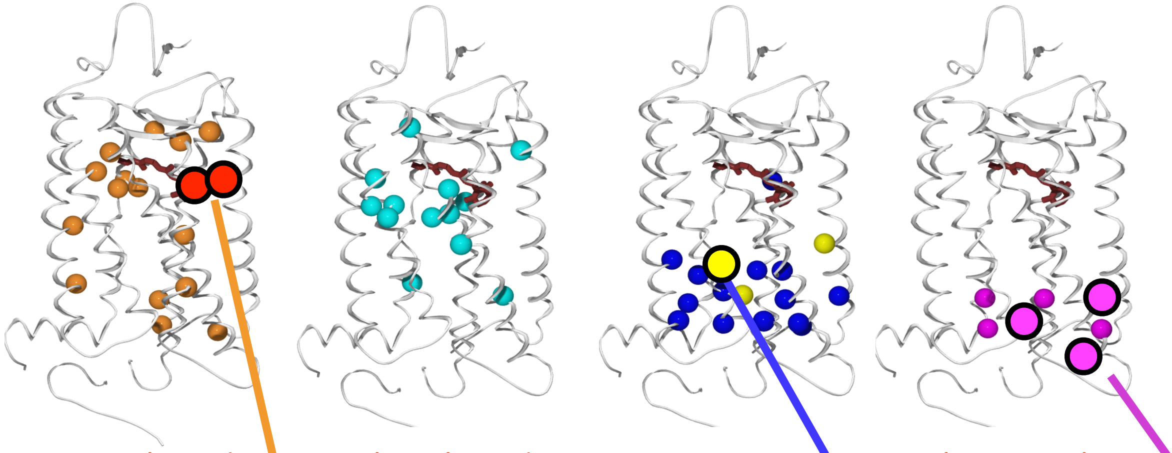
Common G protein activation

RATIONAL RE-DESIGN OF GPCR SIGNALING

Lichtarge Lab
Baylor College of Medicine

SPECIFIC | **GENERIC**

Ligand Binding | Trigger | Conformational Switch | G-protein coupling



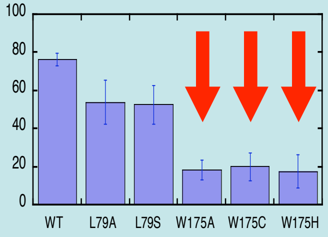
Swap ligands
Target drug

Alter ligand
sensitivity

Constitutive
activity

Alter coupling
specificity

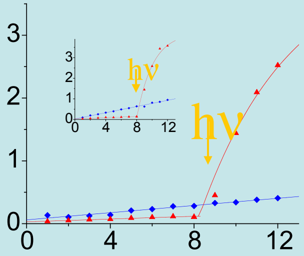
Impaired retinal binding



% Rhodopsin Regeneration

Alecia Gross et al.

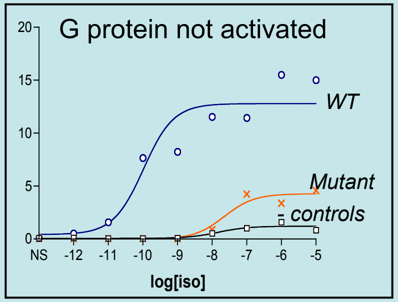
Constitutive activation



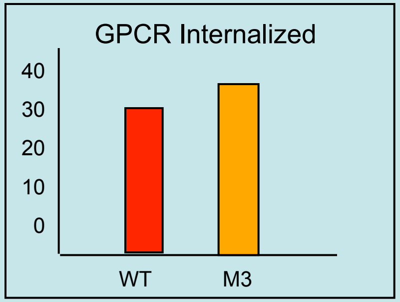
GTPγS bound (pmoles)

Madabushi et al. J. Biol. Chem. 2004

Uncouple efferent pathways



cAMP accumulation HEK-293 (pMoles)

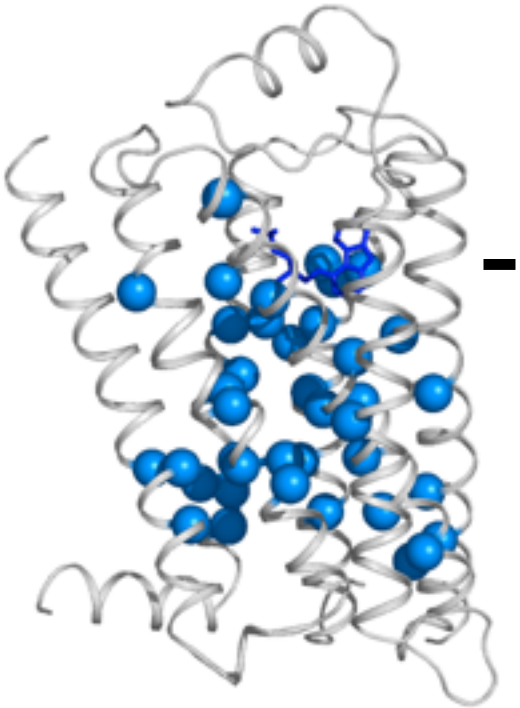


Percent Sequestration β_2-AR mutant

Shenoy et al. J. Biol. Chem. 2006
LICHTARGE/LEFKOWITZ LABS

STRATEGY TO IDENTIFY DETERMINANTS

Gustavo Rodriguez
Ted Wensel

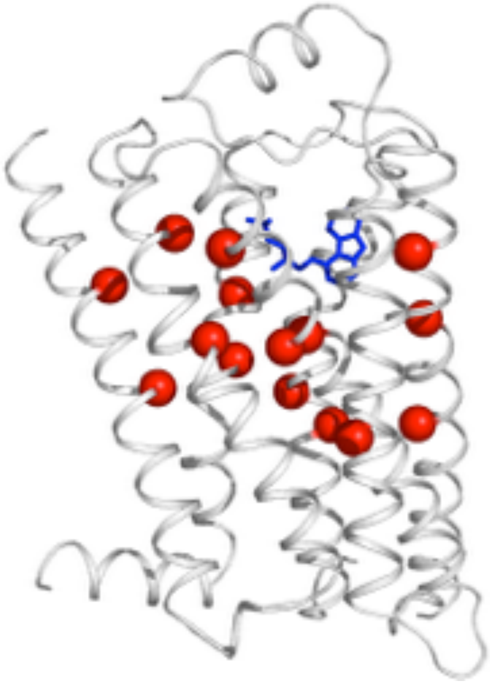


Amine ET

-

Class A

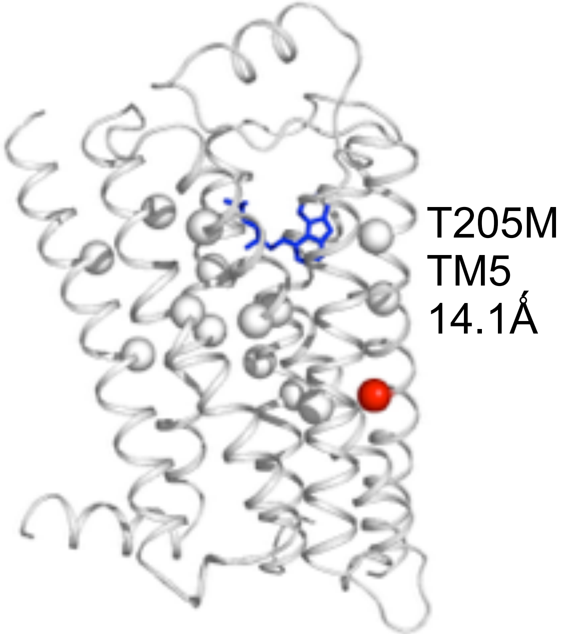
=>



Difference ET

TM5:T205M DECOUPLING AND SWITCHING

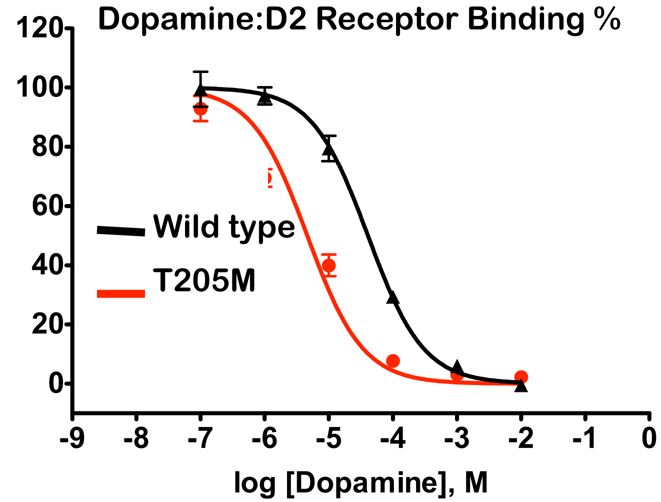
Gustavo Rodriguez
Ted Wensel

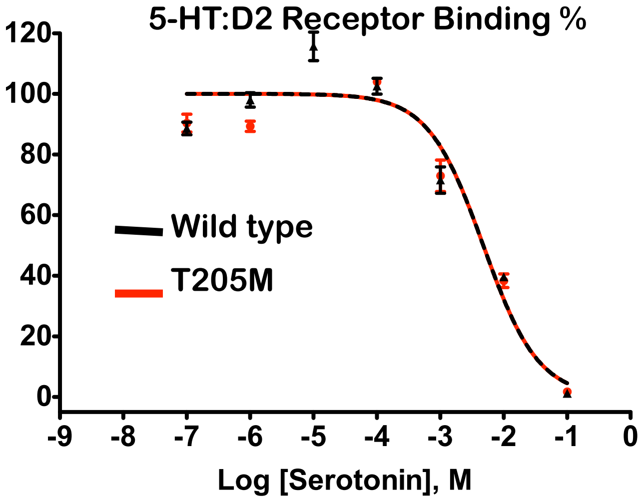


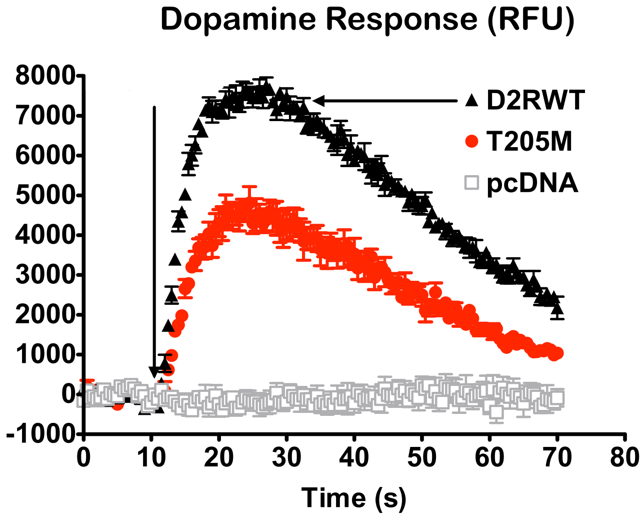
Far

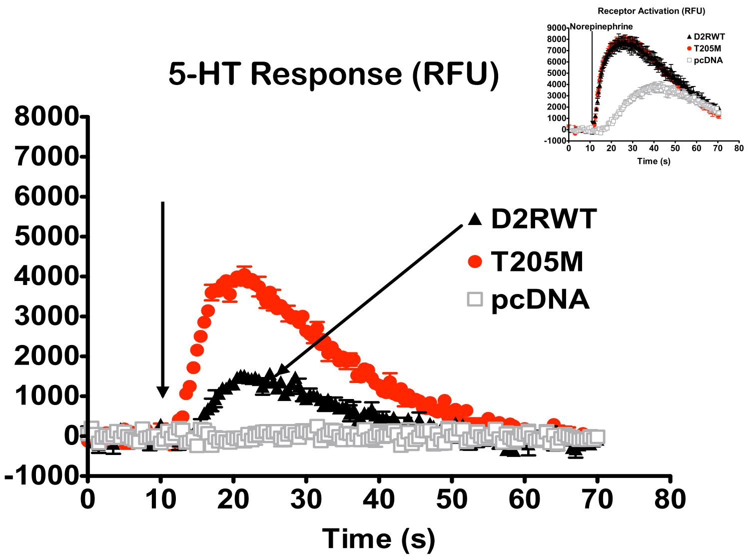
Better dopamine binding

Partial dopa vs 5-HT switch





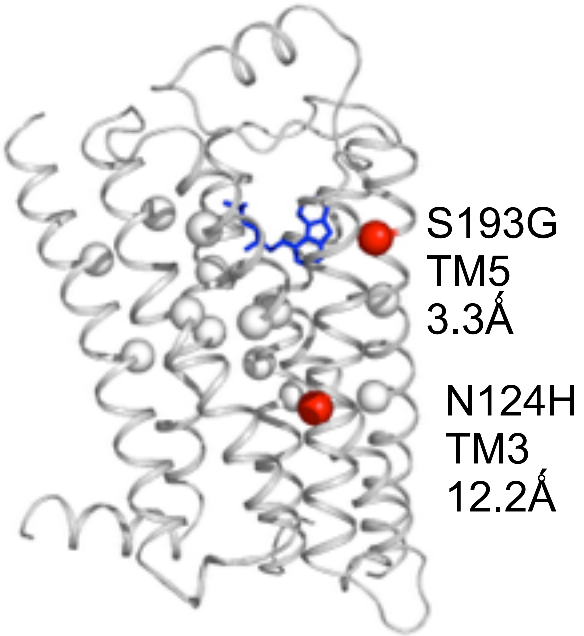


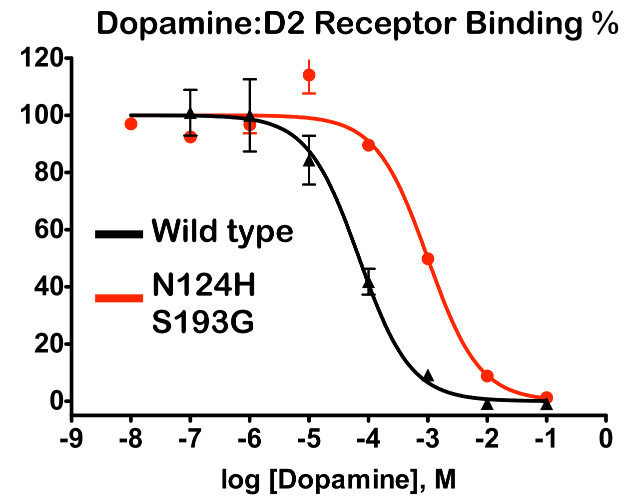


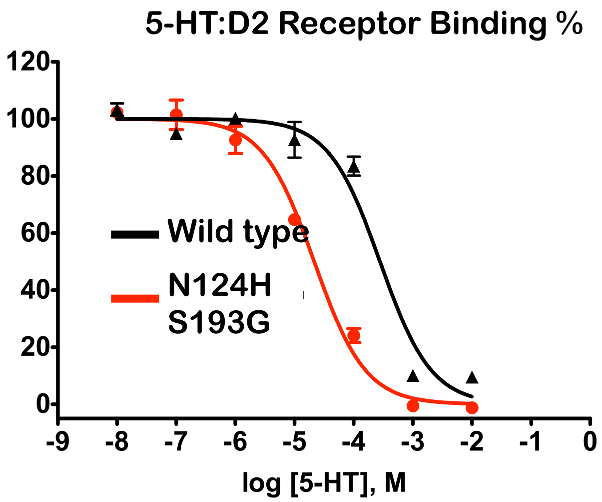
SYNERGY

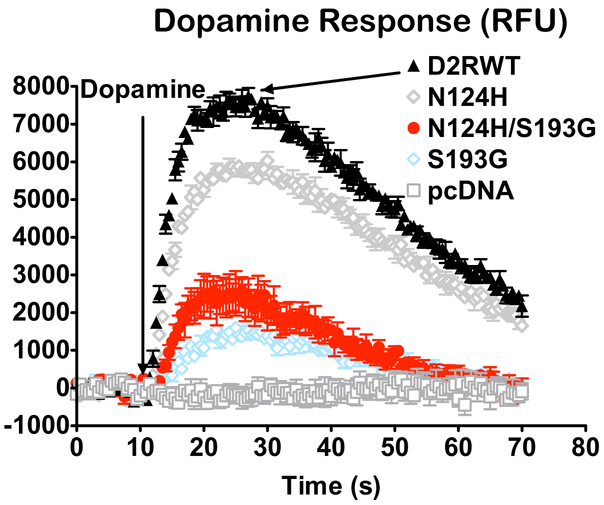
Lichtarge Lab
Baylor
College of
Medicine

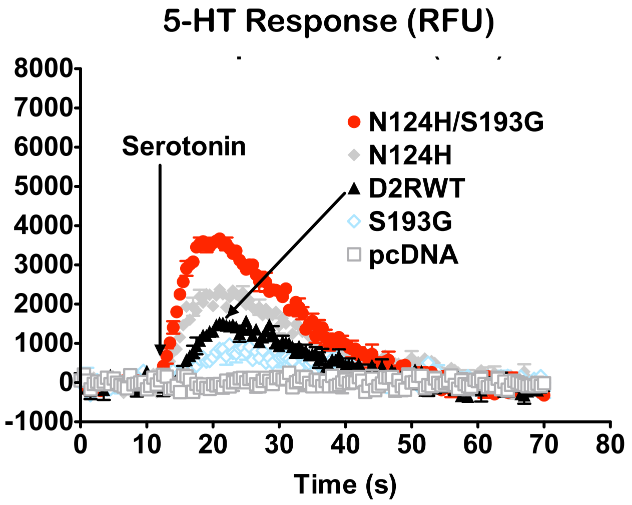
Gustavo Rodriguez
Ted Wensel




Dopamine:D2 Receptor Binding %


5-HT:D2 Receptor Binding %


Dopamine Response (RFU)


5-HT Response (RFU)

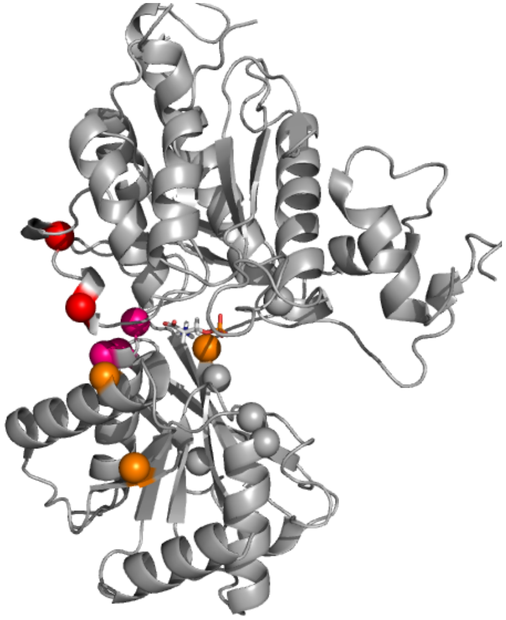
S193G-like binding

<u>*synergistic 5-HT response*</u>

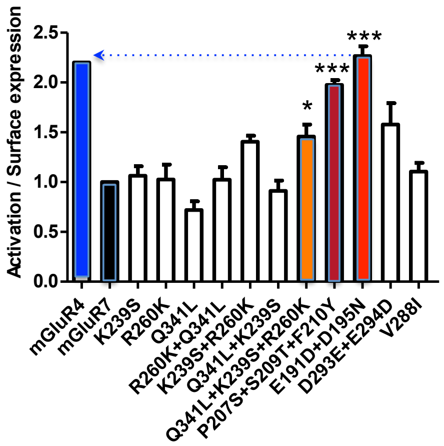
CLASS C GPCR (METABOTROPIC)

Lichtarge Lab
Baylor College of Medicine

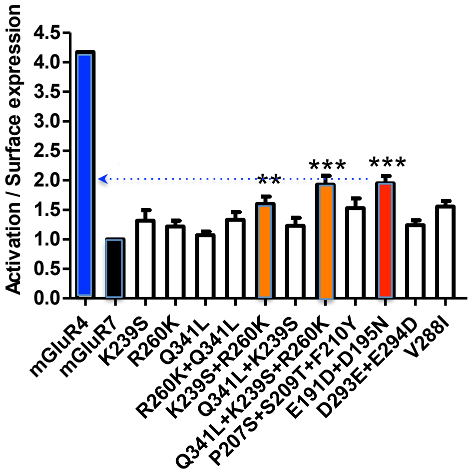
MUTATION SETS



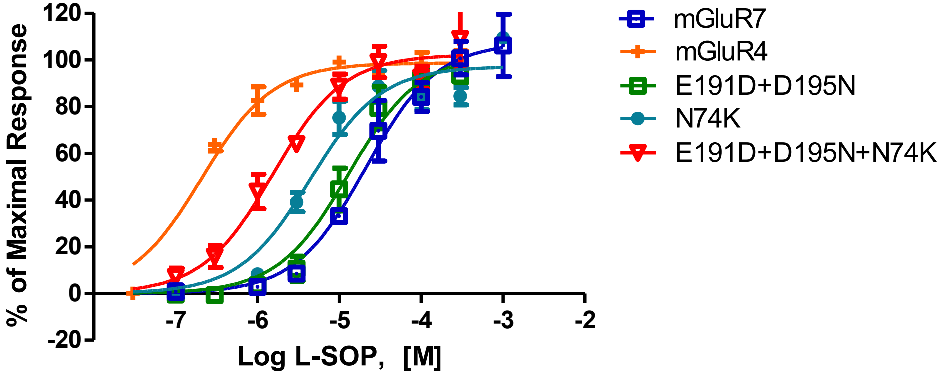
SWITCHING L-SOP EFFICACY



SWITCHING L-GLU EFFICACY



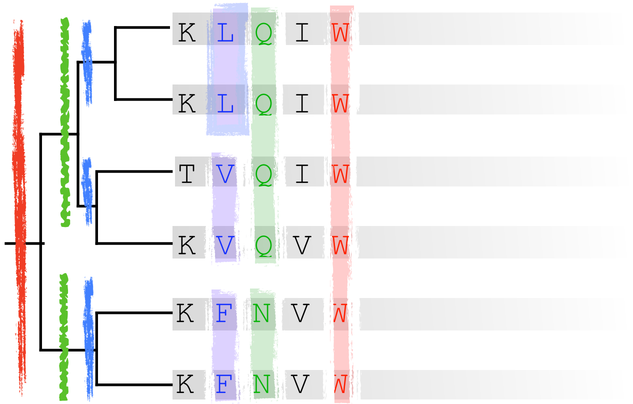
SWITCHING L-SOP POTENCY

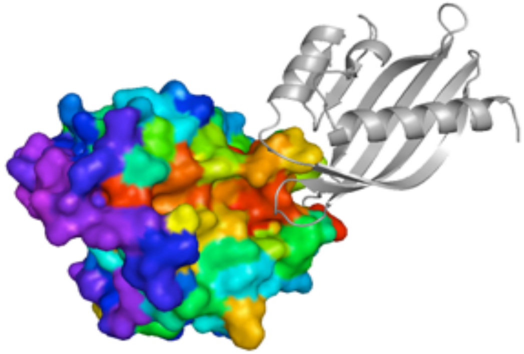


Hye Jin Kang
Ted Wensel

Kang *et al* JBC 2014

RATIONAL RE-DESIGN OF GPCR SIGNALING





Madabushi *J. Biol. Chem* 2004
Shenoy *J. Biol. Chem.* 2006
Bonde *Biochem Pharm* 2010
Koyabashi *Mol. Pharm.* 2010
Rodriguez *PNAS* 2010
Kang et al *JBC* 2014

- **Engineer separation of function**
  - -potency vs efficacy
  - -G protein vs beta-arrestin signaling
  - -L-SOP vs L-GLU
- **Identify site and key residues for**
  - -Ligand binding
  - -G protein coupling
  - -Allosteric pathway
- **Reprogram “allosteric specificity”**
- **Reprogram binding affinity and potency**
- **Functional correlation for some residues**

EXPERIMENTAL CASE STUDIES

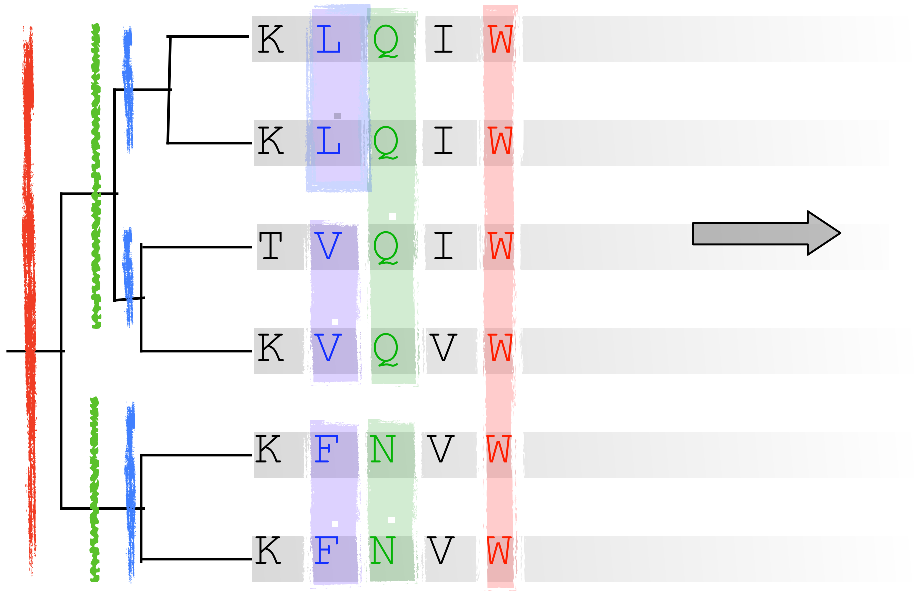
Application/Test	**Protein**	**Collaborator**	**References**	
Predict Binding site	Gα transducin	*Bourne*	*Lichtarge **PNAS** 1996*	*Onrust **Science** 1997*
Separate function	Nuclear Transport	*Moore*	*Cushman **J. Mol. Biol.** 2004*	
	Steroid receptors	*Smith*	*Raviscioni **Proteins** 2006*	
	Ku	*Bertuch*	*Ribes-Zamora **Nat Struc Molec Biol** 2007*	
	RecA - LexA	*Herman*	*Adikesavan **PLoS Genetics** 2011*	
Swap function	RGS	*Wensel*	*Sowa **PNAS** 2000*	*Sowa **Nat Struct Biol** 2001*
	Proneural proteins	*Hassan*	*Quan **Development** 2004*	
	Nuclear Receptors	*Cooney*	*Raviscioni **J. Mol Biol.** 2005*	
Redesign function	GPCRs	*Wensel*	*Yang **Mol Endocrinol** 2002*	*Lin **J Biol Chem**.2003*
		Lefkowitz	*Madabushi **J Biol Chem**.2004*	*Shenoy **J Biol Chem**.2006*
		Bouvier	*Kobayashi **Traffic** 2009*	
		Hansen	*Bonde **Biochem Pharmacol** 2010*	
Build inhibitors	Nuclear Receptors	*Cooney*	*Gu **J. Biol. Chem**.2005*	
	GRK	*Clark*	*Baameur **Mol Pharm** 2010*	
	Cohesin	*Pati*		
Assess gene variations	HPD	*Bodamer*	*Item **Mol Gen Metab** 2007*	
	RRM2B	*Shinawi*	*Shaibani **Arch Neurol.** 2009*	
	TP53	*Myers*		
	CAGI			
Predict protein function	Structural Proteome		*Kristensen **Prot Sci** 2006*	*Kristensen **BMC bioinf** 2008*
			*Ward **PloS One** 2009*	*Ward **Bioinformatics** 2009*
			*Erdin **J Mol Biol** 2010*	*Venner **PLoS One** 2011*

Beyond these case studies, is ET general ?

Scalability of ET

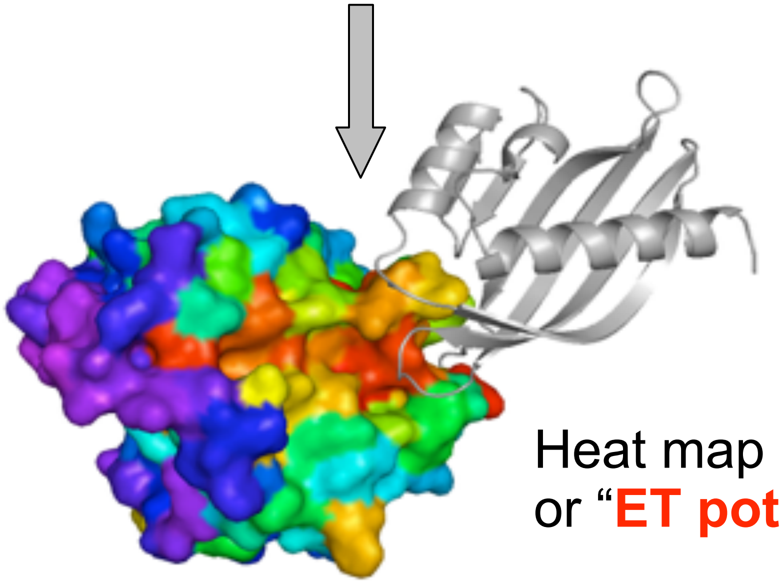
Towards Laws of Protein Sequence-Structure-Function

EVOLUTIONARY TRACE: BUILDING 3D TEMPLATES



*rv*ET

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \{ - \sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \}$$



Heat map of ET rank
or "**ET potential**"

Match evolutionary with sequence *variations*

Big *divergences* suggest important positions

Lichtarge ***J.Mol Biol*** *(1996)*
Lichtarge ***J.Mol Biol*** *(1997)*
Madabushi ***J.Mol Biol*** *(2002)*
Mihalek ***J.Mol Biol*** *(2003)*

Yao & Kristensen ***J.Mol Biol*** *(2003)*
Mihalek ***J.Mol Biol*** *(2004)*
Mihalek ***Proteins*** *(2006)*
Mihalek ***Bioinformatics*** *(2006)*

Yao ***Proteins*** *(2006)*
Morgan ***Bioinformatics*** *(2006)*
Mihalek et al ***J.Mol Biol*** *(2007)*
Mihalek et al. ***BMC bioinformatics*** *(2007)*

Wilkins ***Protein Sci*** *(2010)*
Wilkins ***Bioinformatics*** *(2013)*

SEQUENCE-STRUCTURE-FUNCTION LAWS

Lichtarge Lab
Baylor College of Medicine

Important ET residues cluster

ET clusters are general

ET clusters overlap sites

Significant overlaps are general

Clustering and overlap correlate

The correlation is general-50 proteins

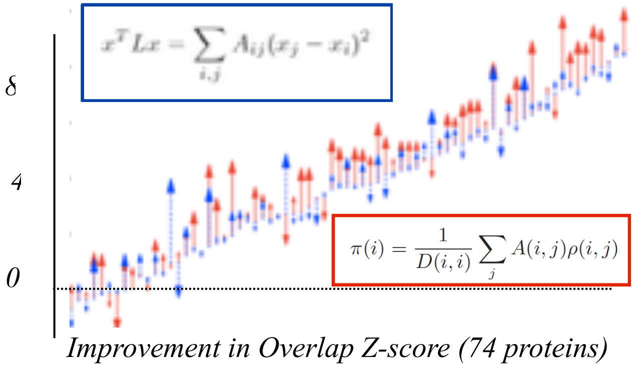
Cluster optimization improves overlap

Most clustering optimizations raise overlap

Smoothing ET ranks maximizes overlap

The improvement in overlap is robust

Smoothing ET ranks improves overlap



$$x^T L x = \sum_{i,j} \overbrace{A(i,j)}^{\text{Structure Information}} \overbrace{(x_j - x_i)}^{\text{Sequence Information}}{}^2$$

To measure smoothness of a feature we describe the structure as a graph.

$$\pi(i) = \overbrace{\frac{1}{D(i,i)} \sum_j A(i,j)}^{\text{Structure Information}} \overbrace{\rho(i,j)}^{\text{Sequence Information}}$$

Test Set (74 proteins)

	rvET	piET
$\langle x^t L x \rangle$	0.214	0.127
$\langle z_c \rangle$	6.26	9.91
$\langle z_o \rangle$	3.71	4.55

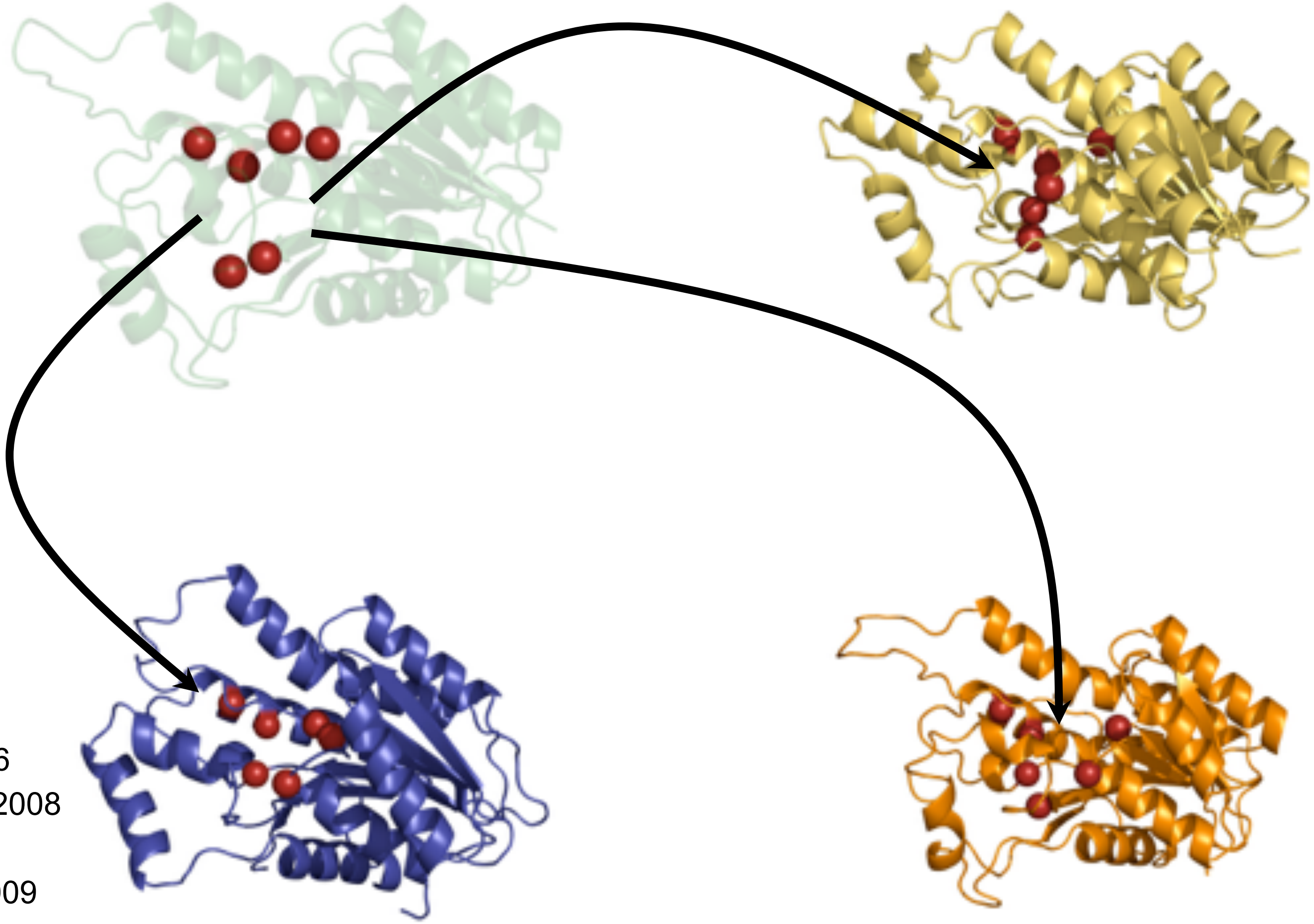
*Lichtarge **J.Mol Biol** (1996)*
*Lichtarge **J.Mol Biol** (1997)*
*Madabushi **J.Mol Biol** (2002)*
*Mihalek **J.Mol Biol** (2003)*

*Yao & Kristensen **J.Mol Biol** (2003)*
*Mihalek **J.Mol Biol** (2004)*
*Mihalek **Proteins** (2006)*
*Mihalek **Bioinformatics** (2006)*

*Yao **Proteins** (2006)*
*Morgan **Bioinformatics** (2006)*
*Mihalek et al **J.Mol Biol** (2007)*
*Mihalek et al. **BMC bioinformatics** (2007)*

*Wilkins **Protein Sci** (2010)*
*Wilkins **Bioinformatics** (2013)*

STRUCTURAL PROTEOMIC ANNOTATION OF FUNCTION



Kristensen ***Prot Sci*** 2006

Kristensen ***BMC Bioinf*** 2008

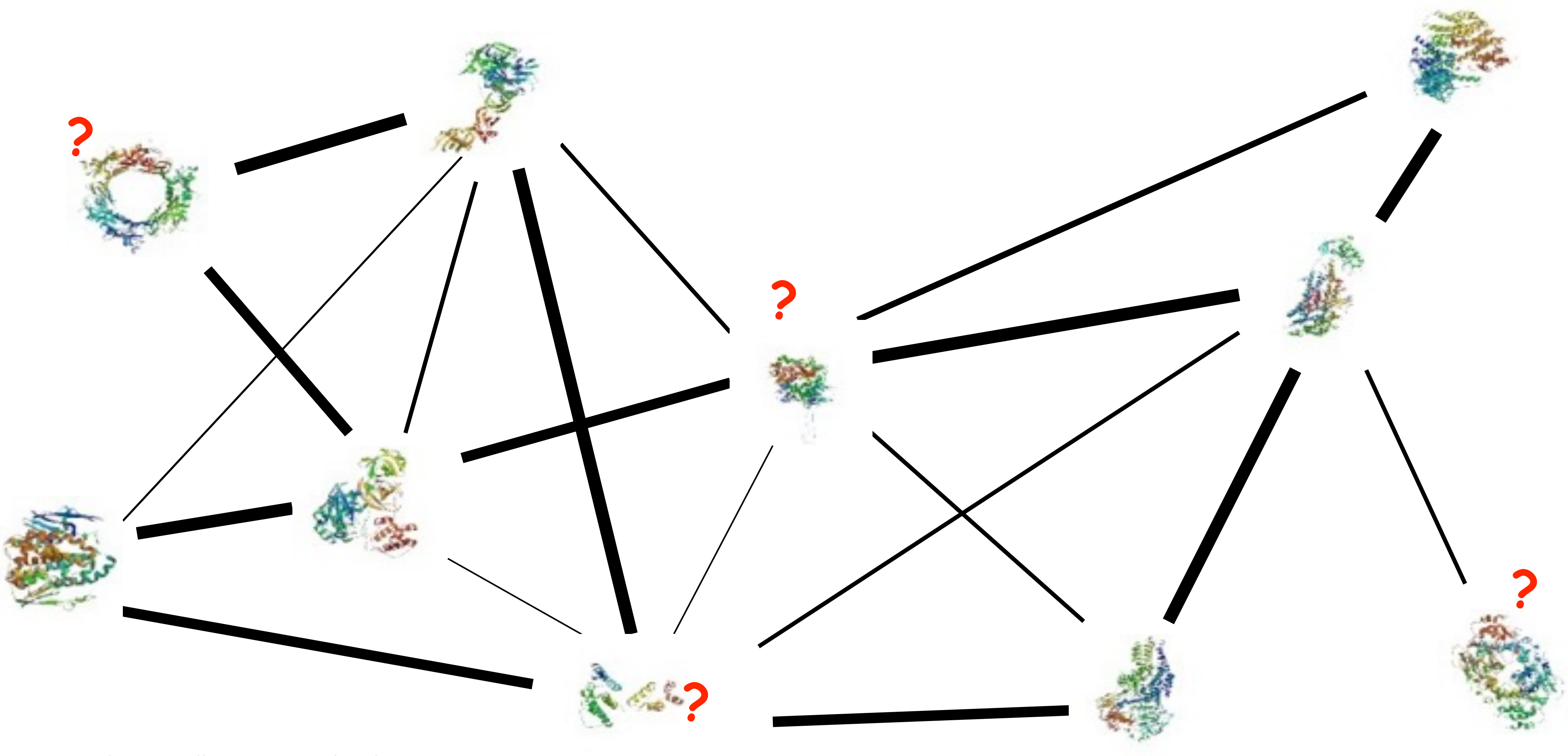
Ward ***PloS One*** 2009

Ward ***Bioinformatics*** 2009

Erdin ***J Mol Biol*** 2010

Venner and Lisewski et al ***PLoS ONE*** 2011
Bachman et al ***Bioinformatics*** 2012
Amin *et al* ***PNAS*** 2013

DIFFUSE AND INTEGRATE FUNCTIONAL TAGS

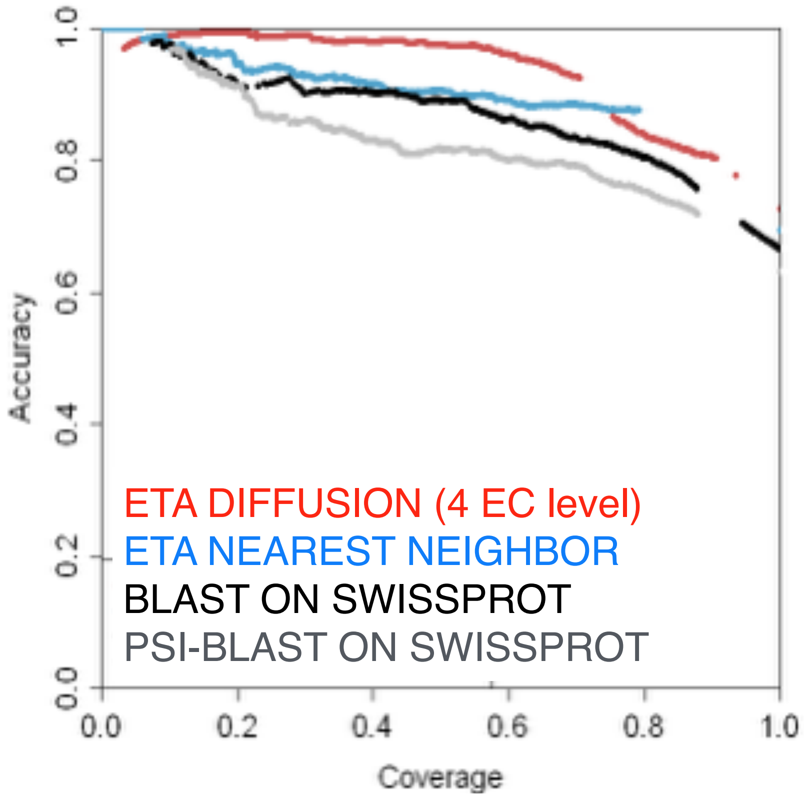


Images from http://www.rcsb.org/pdb/

SMOOTHNESS OVER NETWORK STRUCTURES

ANNOTATIONS OF 1217 SGI PROTEIN CONTROLS

ETA IMPROVES ON SEQUENCE ANNOTATION

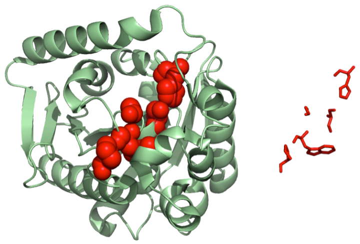


Venner *et al* ***PLoS One*** 2010
Bachman *et al* ***Bioinformatics*** 2012

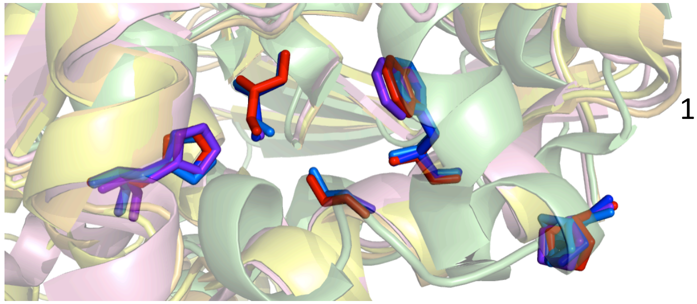
FUNCTION PREDICTION (ON STRUCTURE)

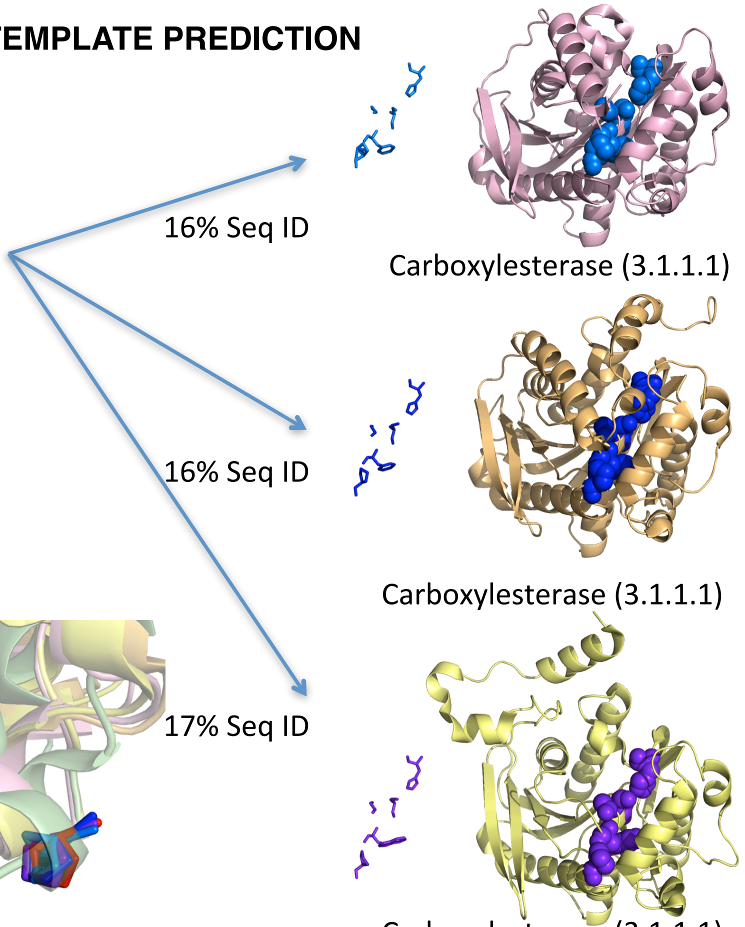
Lichtarge Lab
Baylor College of Medicine

CASE STUDY OF A 5-RESIDUE TEMPLATE PREDICTION



Putative	E.C.
Esterase	(3.1.x.x)
Lipase	(3.1.1.3)
Thioesterase	(3.1.2.x)





16% Seq ID

Carboxylesterase (3.1.1.1)

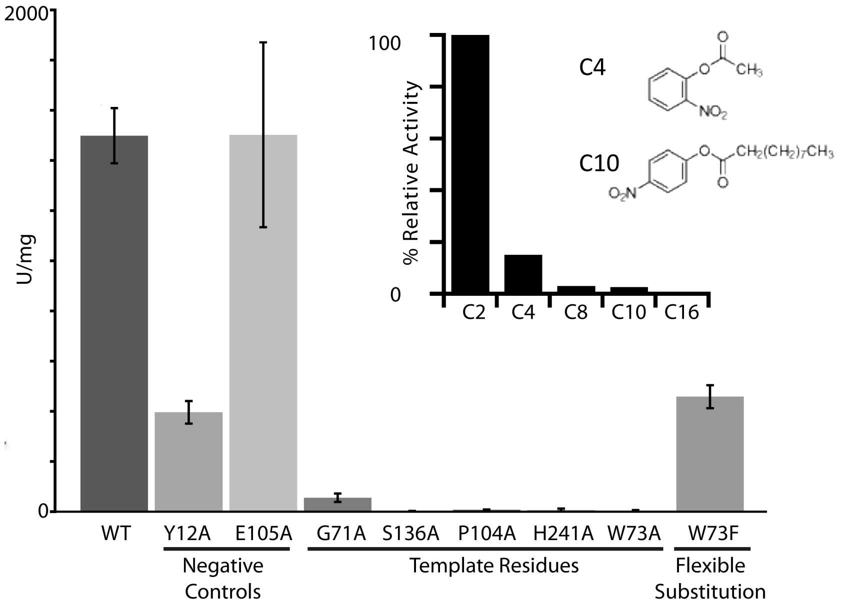
16% Seq ID

Carboxylesterase (3.1.1.1)

17% Seq ID

Carboxylesterase (3.1.1.1)

A TRUE SMALL FATTY ACID CARBOXYL ESTERASE

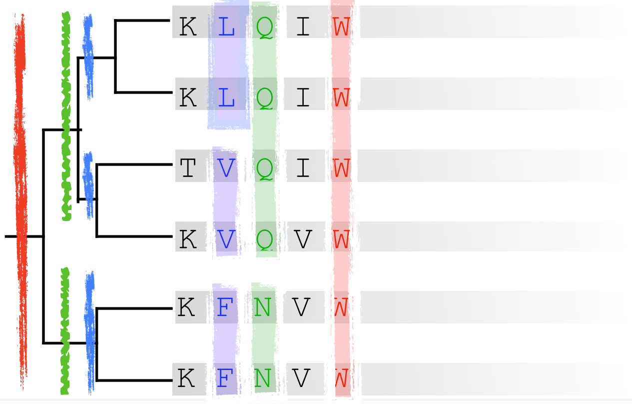


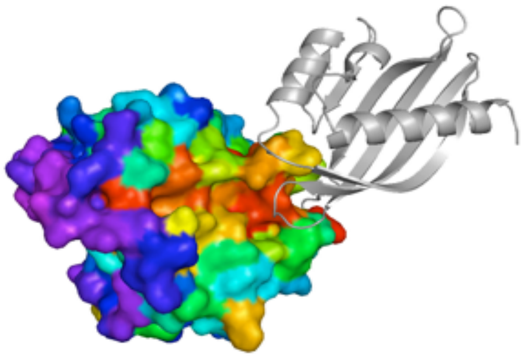
These data show:

- High resolution functional information mined from structure and evolution
- Small ET templates can be highly specific and accurate deep in twilight zone
- Other studies further network these templates (Venner et al PLoS One 2010)

Amin *et al* ***PNAS*** 2013

VARIATIONAL ASPECTS OF ET





Couple sequence variations with evolutionary variations

Clusters of top ET residues map functional determinants and sites

- Proteome scale
- High resolution (substrates)
- Deep in “twilight” zone

An extremum of ET rank variations optimizes information recovery

- Improves with smoothing ET
- Improves with network diffusion

Venner and Lisewski et al ***PLoS ONE*** (2010)
Bachman et al ***Bioinformatics*** (2012)
Amin and Erdin et al ***PNAS*** (2013)
Wilkins ***Bioinformatics*** *(2013)*

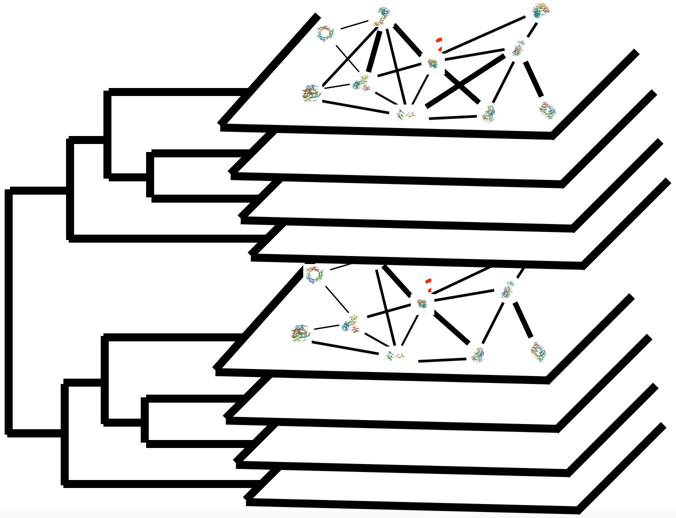
NETWORKS AND SMOOTHING

Function prediction without structure

FUNCTION PREDICTION ON BIOMIC NETWORKS

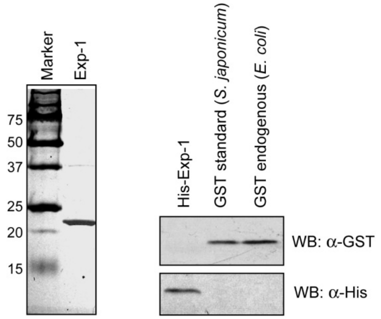
Lichtarge Lab
Baylor College of Medicine

1. **BUILD A BIOMIC NETWORK (373 SPECIES)**
2. **COMPRESS IT BY COLLAPSING CLIQUES (COGS)**
3. **DIFFUSE INFORMATION <u>SMOOTHLY</u> (IN TIME/STRUCTURE)**
4. **EXP-1: A GLUTATHIONE-S-TRANFERASE (9% IDENTITY TO HOMOLOG) WITH HEMATIN SUBSTRATE**
5. **ARTESUNATE, A HERBAL DERIVATIVE OF UNKNOWN MECHANISM, INHIBITS EXP-1**

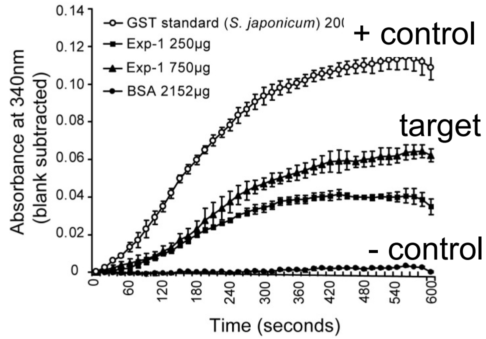


THE FIRST VALIDATED TARGET FOR THE MOST EFFECTIVE ANTI-MALARIAL DRUG:
A path to new drugs?

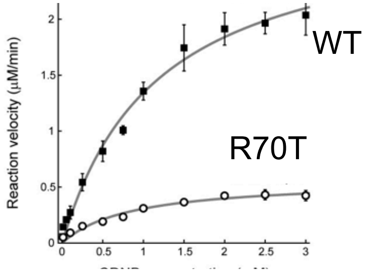
EXP-1 EXPRESSION



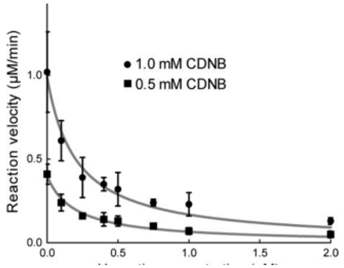
EXP-1 HAS GST ACTIVITY



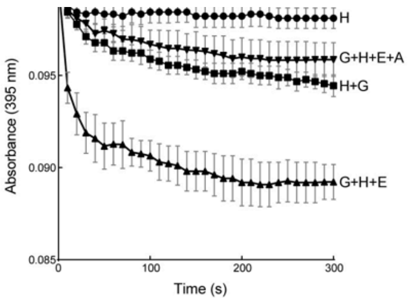
A MUTATION SLOWS REACTION VELOCITY



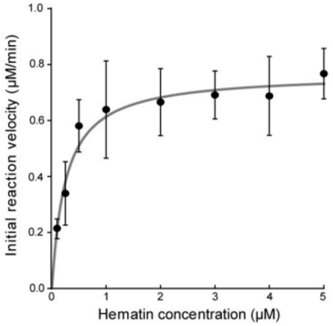
HEMATIN IS AN EXP-1 SUBSTRATE



ARTESUNATE INHIBITS EXP-1



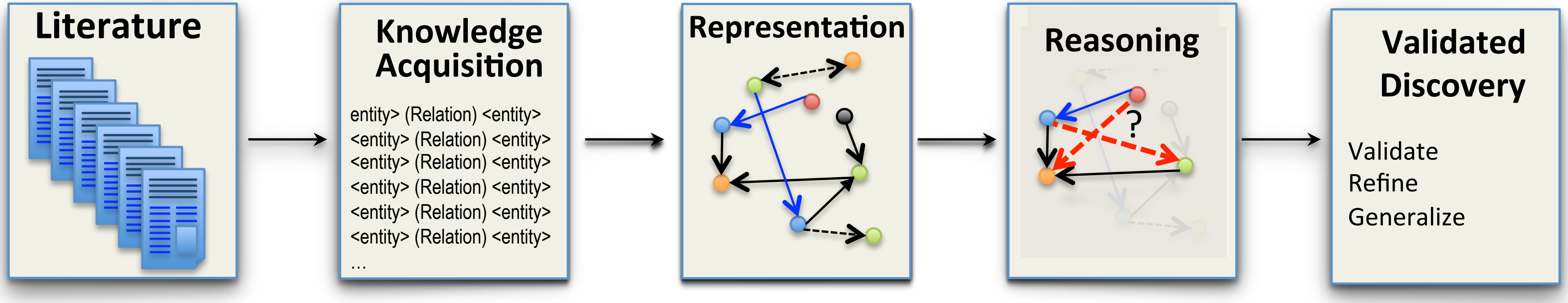
MICHAELIS-MENTEN FIT



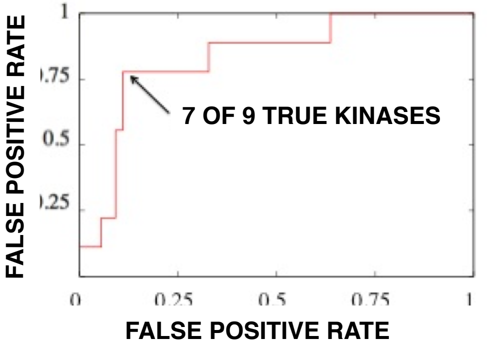
Lisewski and Quiros et al (2014) Cell

AUTOMATED HYPOTHESIS GENERATION FROM THE SCIENTIFIC LITERATURE

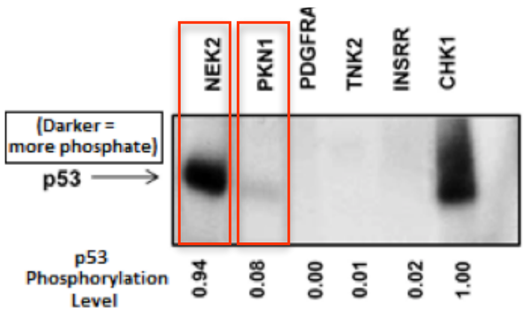
Lichtarge Lab
Baylor College of Medicine



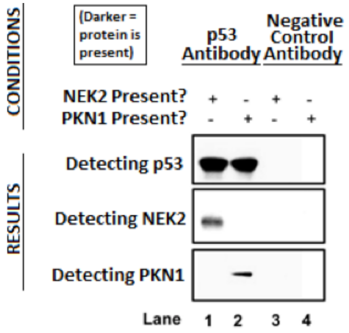
p53 KINASES DISCOVERED POST-2003 BASED ON PRE-2003 LITERATURE



TWO PREDICTED p53 KINASES PHOSPHORYLATE p53



TWO PREDICTED p53 KINASES PHOSPHORYLATE p53



Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014)

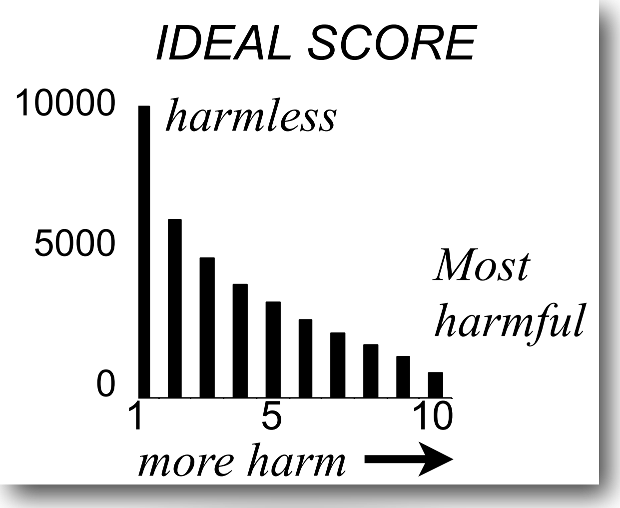
Scott Spangler*[1], Angela D. Wilkins*[3], Benjamin J. Bachman[3], Meena Nagarajan[1],
Tajhal Dayaram[3], Peter Haas[1], Sam Regenbogen[3], Curtis R. Pickering[2], Austin Comer[2],
Jeffrey N. Myers[2], Ioana Stanoi[1], Linda Kato[1], Ana Lelescu[1], Jacques J. Labrie[1],
Neha Parikh[3], Andreas Martin Lisewski[3], Lawrence Donehower[3], Ying Chen[1], Olivier Lichtarge[3]

Part 2. THE CALCULUS OF EVOLUTION

Mutational Impact
in
Proteins,
Diseases,
and Populations

CURRENT MODELS DO NOT YIELD THE EXPECTED DISTRIBUTION OF HARMFUL MUTATIONS

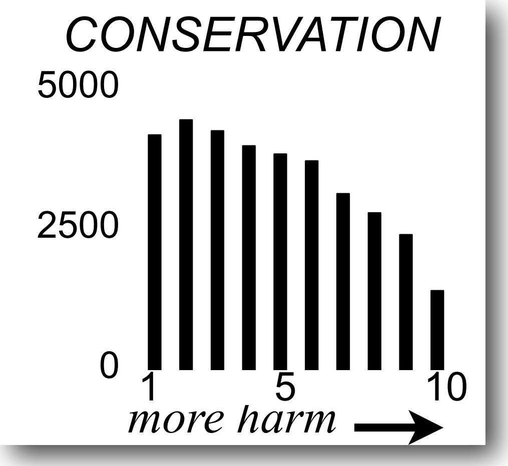
Lichtarge Lab
Baylor College of Medicine

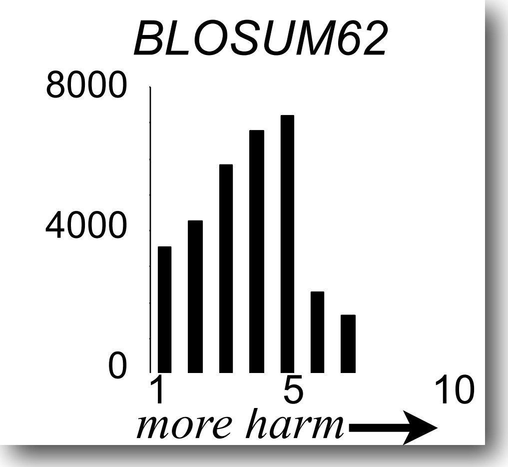


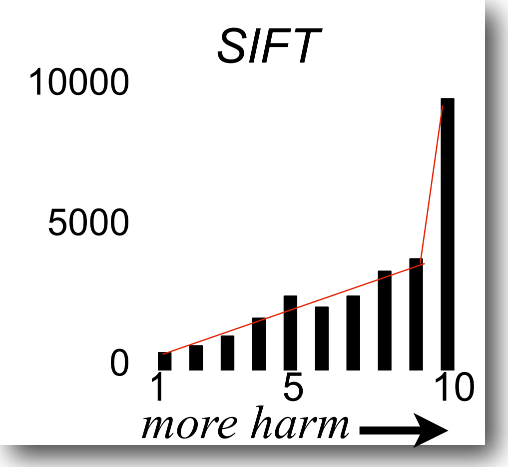
$$frequency \propto e^{-harm}$$

Fisher 1930

Human coding polymorphisms



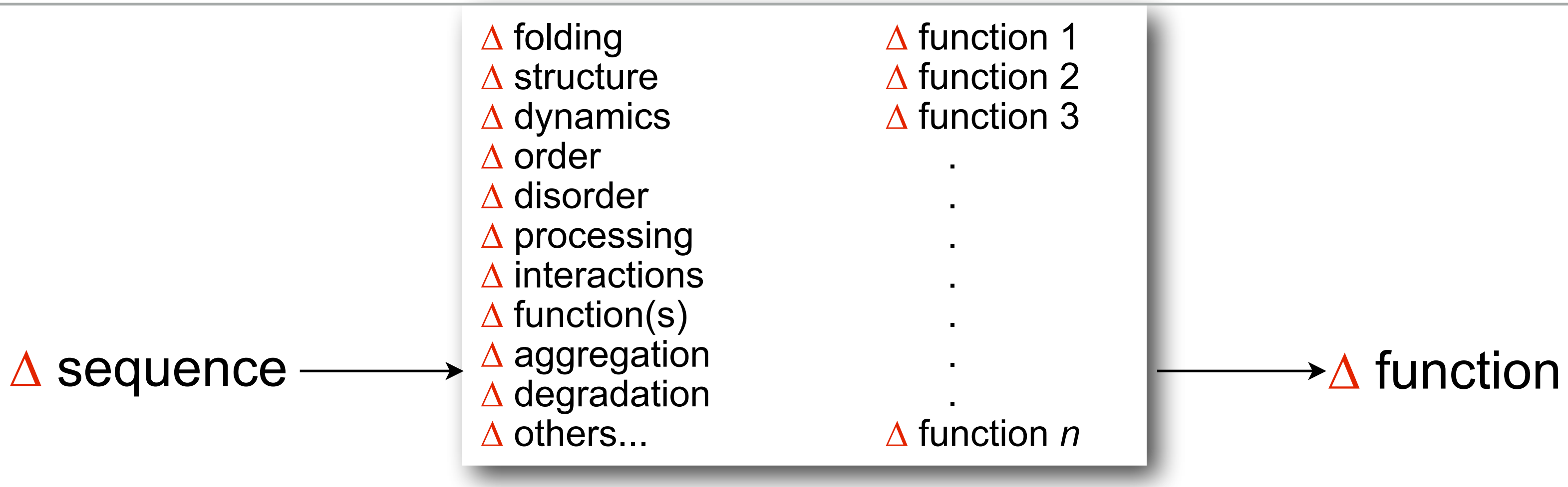




3

Evolutionary Action (EA)

AN EXACT ANALYSIS SEEMS FUTILE



Δ sequence →

- Δ folding
- Δ structure
- Δ dynamics
- Δ order
- Δ disorder
- Δ processing
- Δ interactions
- Δ function(s)
- Δ aggregation
- Δ degradation
- Δ others...

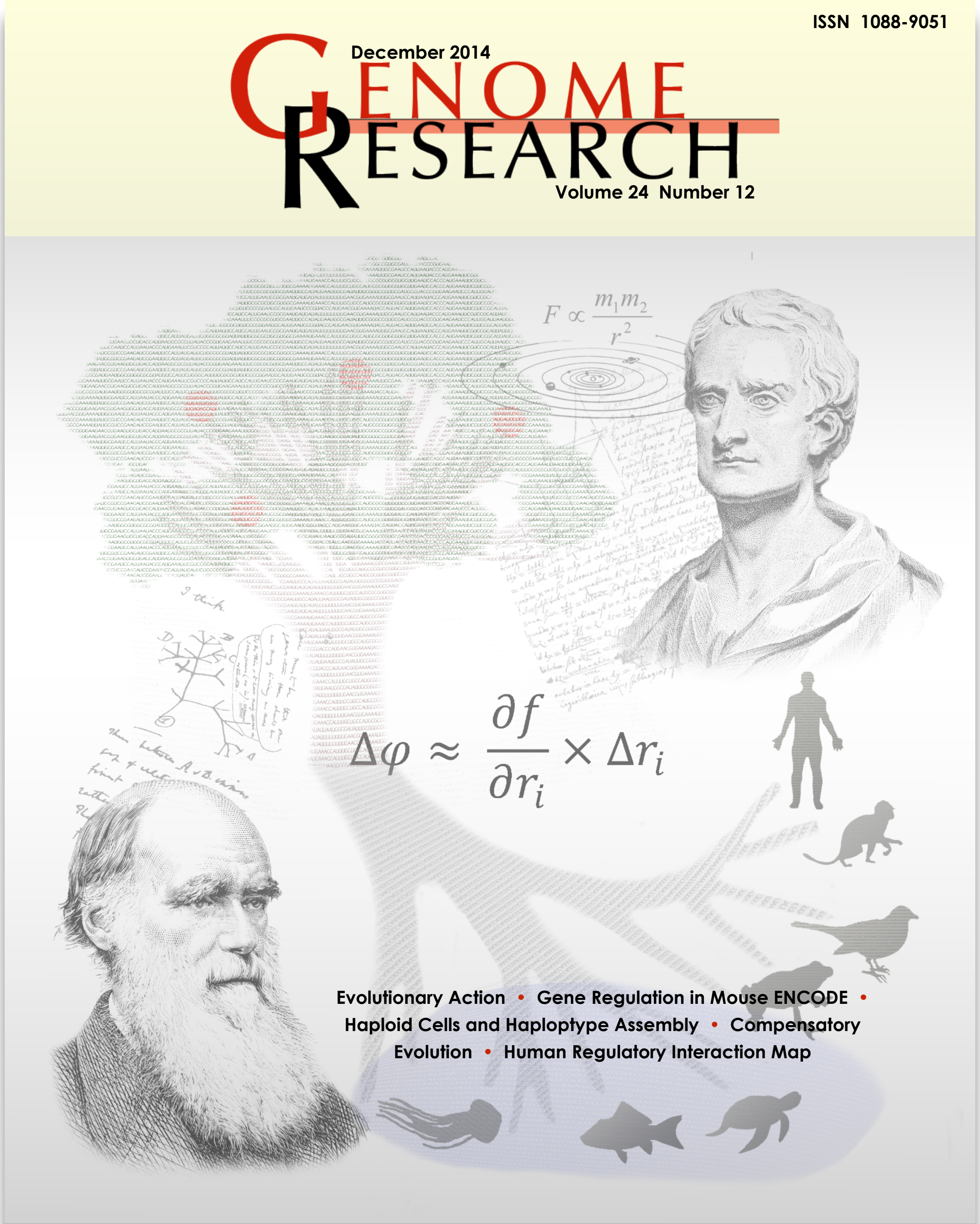
- Δ function 1
- Δ function 2
- Δ function 3

.
.
.
.
.
.
.

- Δ function *n*

→ Δ function

Lichtarge Lab
Baylor
College of
Medicine



ISSN 1088-9051

December 2014

GENOME RESEARCH

Volume 24 Number 12

$$F \propto \frac{m_1 m_2}{r^2}$$

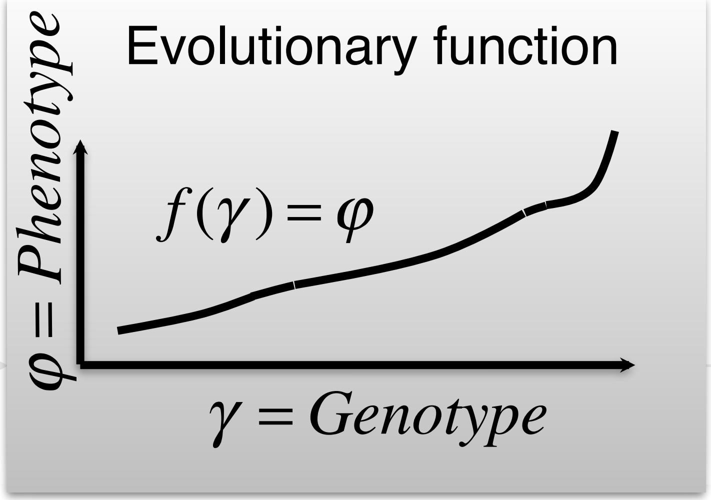
$$\Delta\varphi \approx \frac{\partial f}{\partial r_i} \times \Delta r_i$$

Evolutionary Action • Gene Regulation in Mouse ENCODE • Haploid Cells and Haploptype Assembly • Compensatory Evolution • Human Regulatory Interaction Map

Katsonis and Lichtarge
Genome Research (2014)

PROBLEM SOLVED… IN THEORY

Δ sequence →



→ Δ function

$$\boxed{f(\gamma)=\varphi}$$

$$\begin{tikzcd}
\Delta \text{ sequence} \arrow[d] & & \Delta \text{ function} \arrow[d] \\
d\gamma \arrow[r] & \boxed{f'(\gamma)\bullet d\gamma = d\varphi} \arrow[r] & d\varphi
\end{tikzcd}$$

Where:

γ the genotype
φ the phenotype
f the evolutionary function

$d\gamma$ the genotype perturbation (mutation)
$d\varphi$ the action of $d\gamma$ on the phenotype/fitness
f' the evolutionary gradient

Katsonis and Lichtarge
Genome Research (2014)

THE ACTION EQUATION

Evolutionary gradient

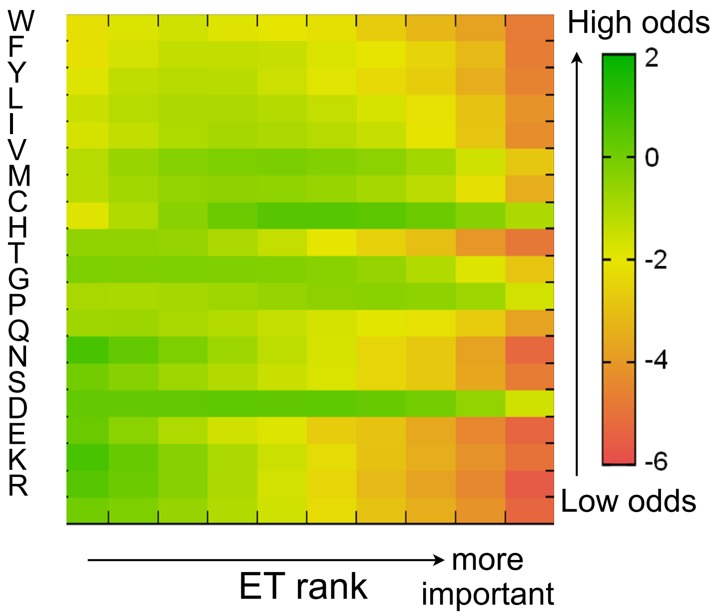
mutation size

$$\boxed{f'(\gamma)\bullet d\gamma = d\varphi}$$

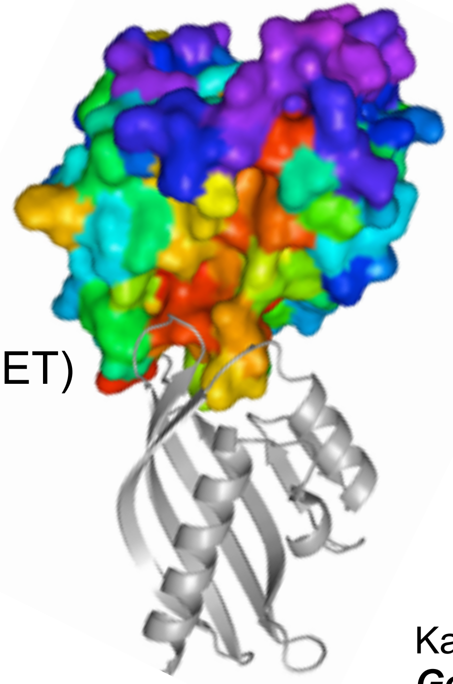
Action of mutation on function or fitness

Phylogenetic variation associated with a sequence variation

1/Substitution Odds

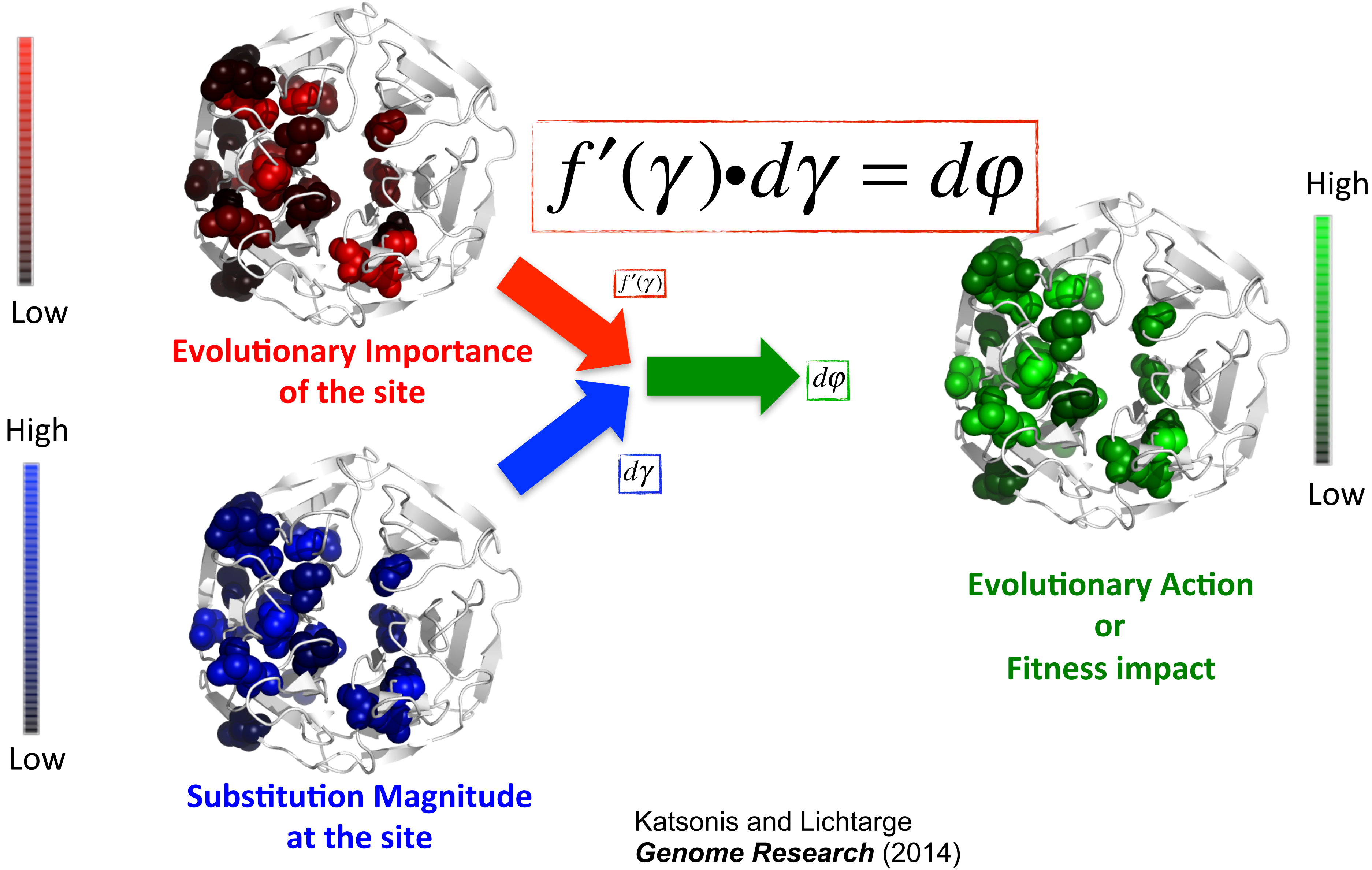


Evolutionary Trace (ET)



Katsonis and Lichtarge
Genome Research (2014)

SO THE ACTION EQUATION CAN BE SOLVED FOR MOST MUTATIONS IN MOST PROTEINS IN MOST SPECIES



$$f'(\gamma)\cdot d\gamma = d\varphi$$

Evolutionary Importance of the site

Substitution Magnitude at the site

Evolutionary Action or Fitness impact

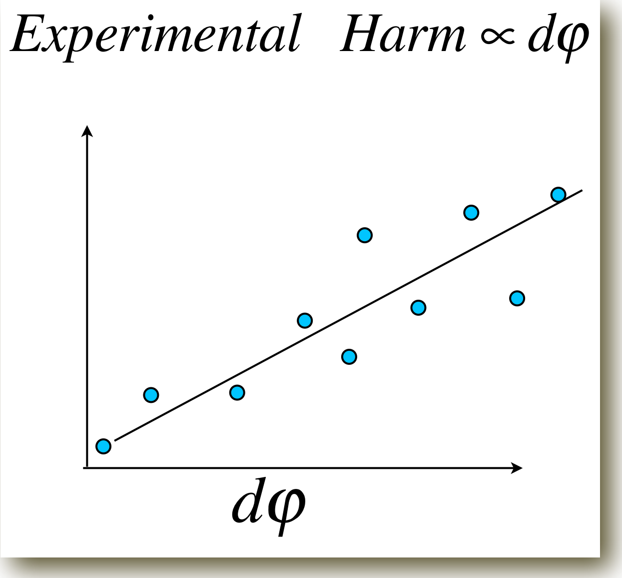
Katsonis and Lichtarge
Genome Research (2014)

Validation

EVOLUTIONARY ACTION (EA) CORRELATES WITH FUNCTIONAL LOSS

$$\boxed{f'(\gamma)\bullet d\gamma = d\varphi}$$

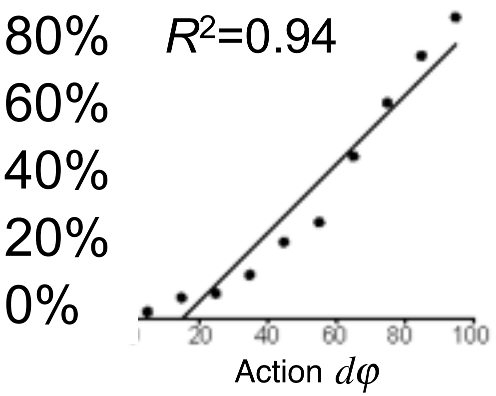
Validation 1



EVOLUTIONARY ACTION (EA) CORRELATES WITH FUNCTIONAL LOSS

Lac Repressor

Loss of β-galactosidase repression
in 4041 mutants

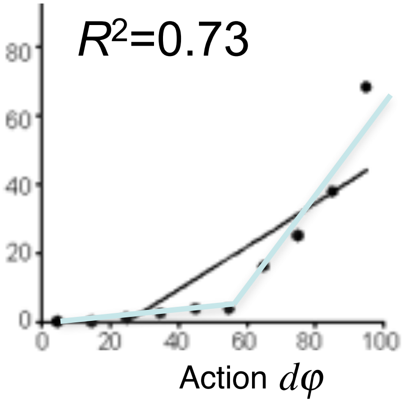


in vivo

bacterial

T4 Lyzozyme

Loss of plaques
in 2015 mutants

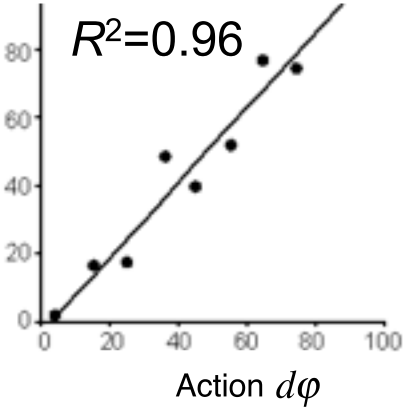


in vivo

bacterial

HIV Protease

loss of cleavage
in 336 mutants

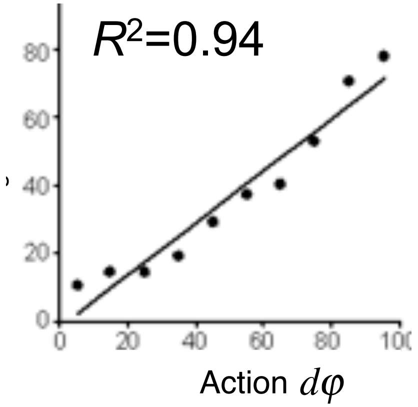


in vitro

viral

p53

mean loss of transactivation
in 2314 mutants



in vitro

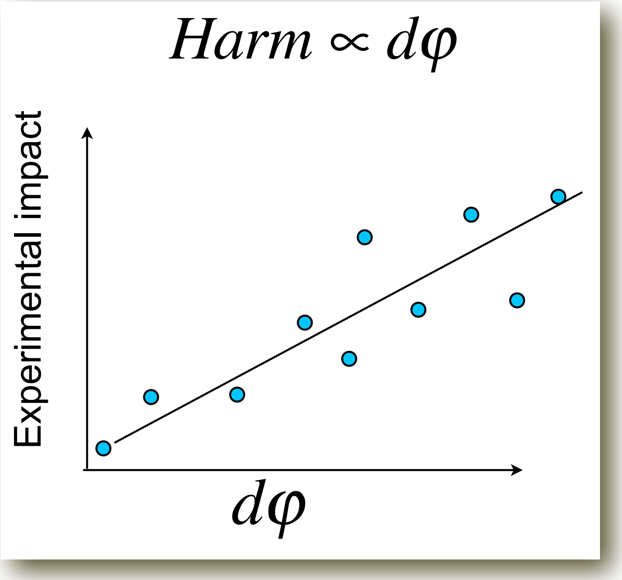
eukaryotic

Katsonis and Lichtarge
Genome Research (2014)

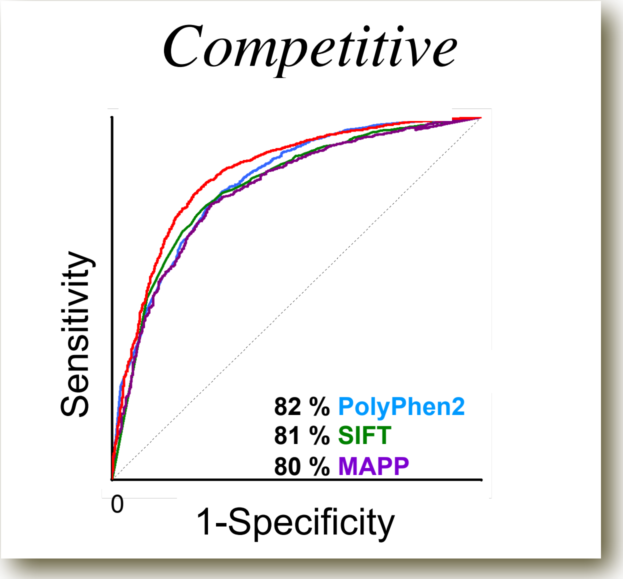
EVOLUTIONARY ACTION (EA) VS OTHER METHODS

$$\boxed{f'(\gamma)\bullet d\gamma = d\varphi}$$

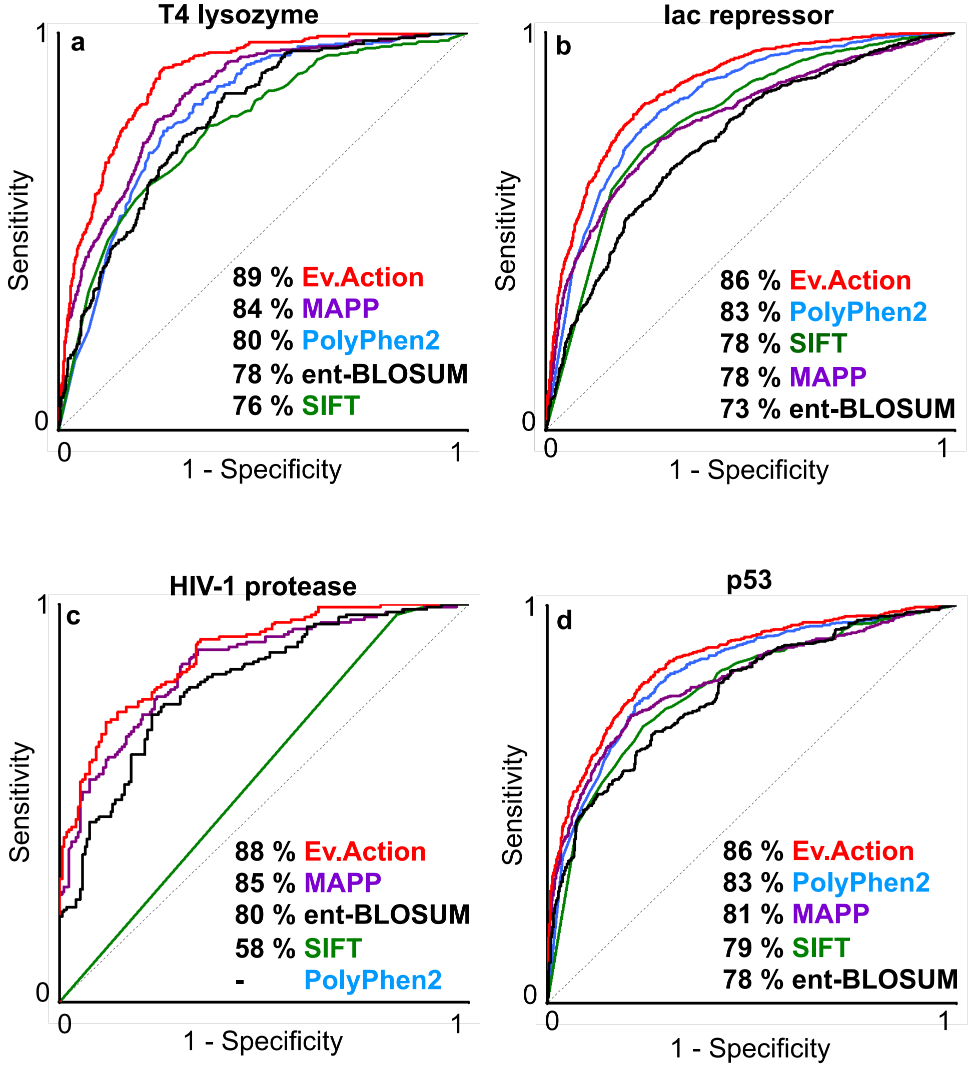
Validation 1 √



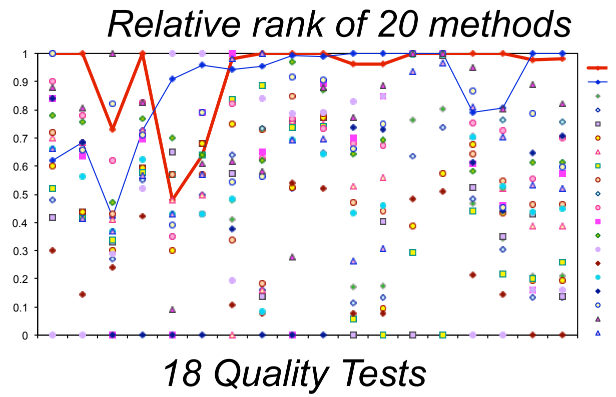
Validation 2

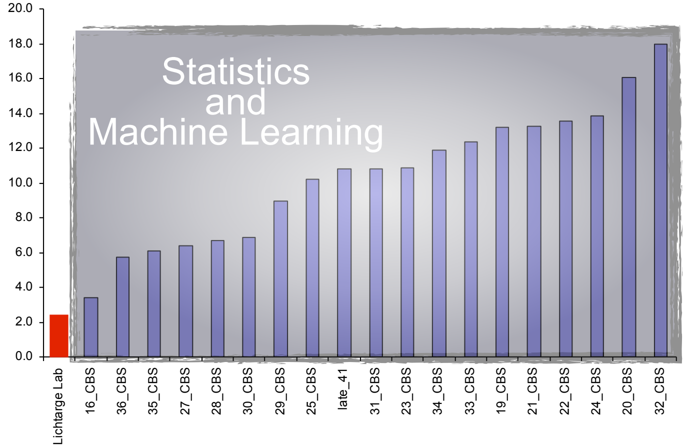


EVOLUTIONARY ACTION (EA) VS OTHER METHODS



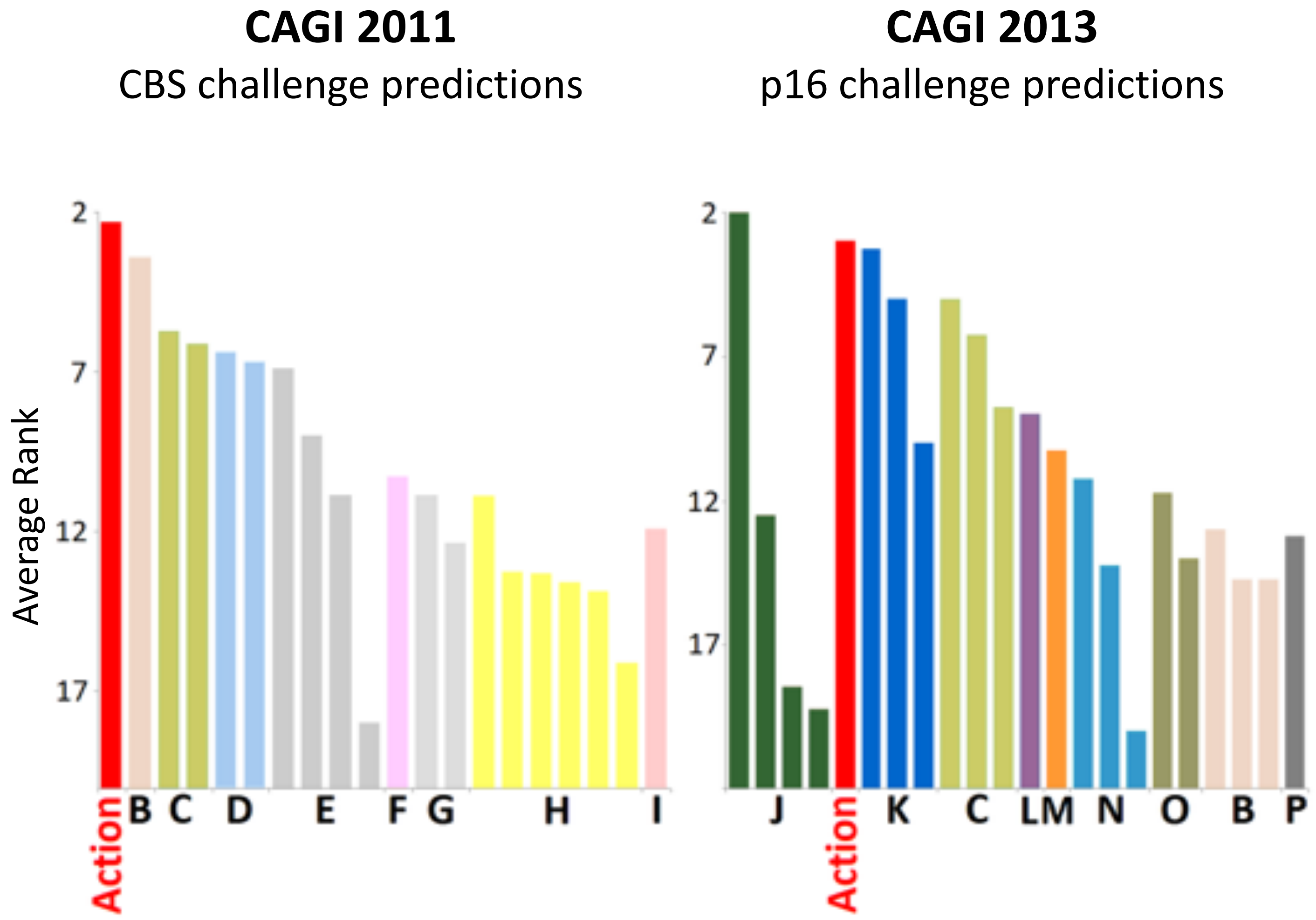
EFFECT OF 84 MUTATIONS OF CYSTATHIONE BETA-SYNTHASE (HOMOCYSTINURIA) ON YEAST GROWTH





Brenner, Repo, Moult

EVOLUTIONARY ACTION (EA) VS OTHER METHODS



CAGI 2011
CBS challenge predictions

CAGI 2013
p16 challenge predictions

PREDICTION METHODS (COLORS INDICATE LAB OF ORIGIN)

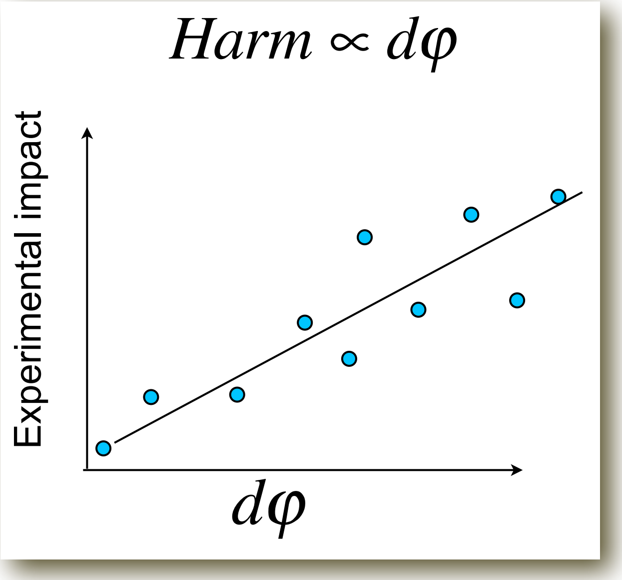
Katsonis and Lichtarge
Genome Research (2014)

EVOLUTIONARY ACTION (EA) RECOVERS FISHER'S DISTRIBUTION

Lichtarge Lab
Baylor College of Medicine

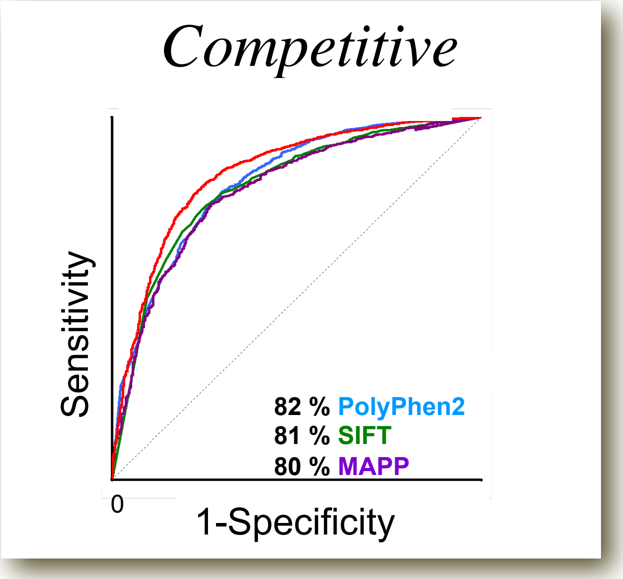
$$\boxed{f'(\gamma)\bullet d\gamma = d\varphi}$$

Validation 1 √



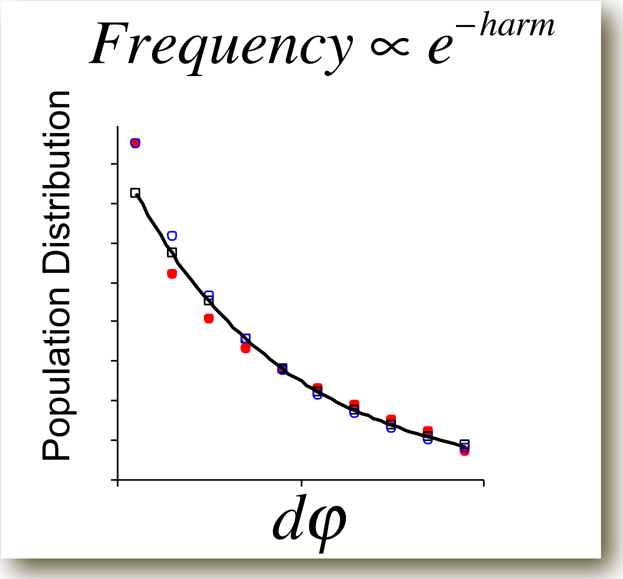
$Harm \propto d\varphi$

Validation 2 √



Competitive

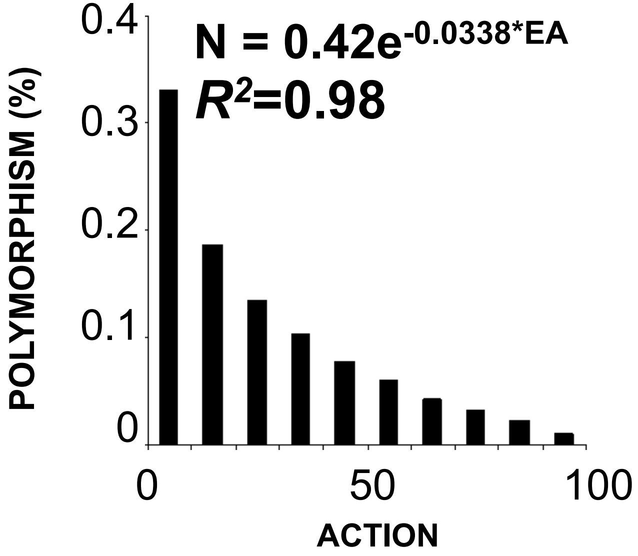
Validation 3



$Frequency \propto e^{-harm}$

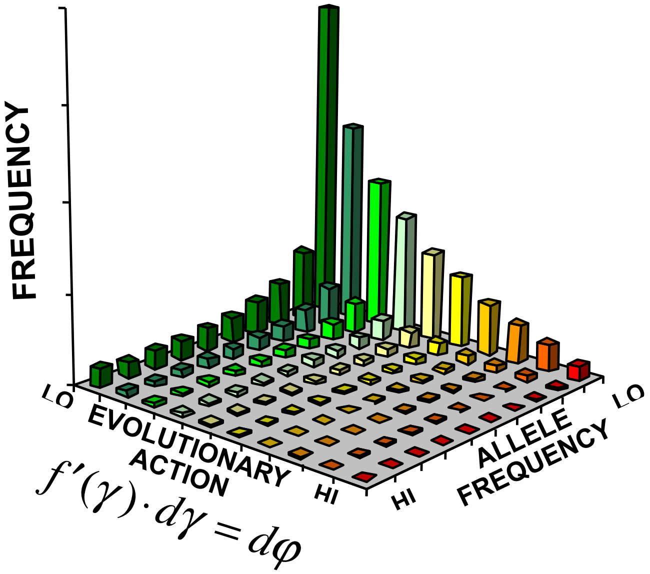
CODING POLYMORPHISMS FROM THE 1000 GENOMES PROJECT DISTRIBUTE EXPONENTIALLY WITH RESPECT TO THEIR ACTION

ACTION DISTRIBUTION OF HUMAN SNPS



$$f'(\gamma)\cdot d\gamma = d\varphi$$

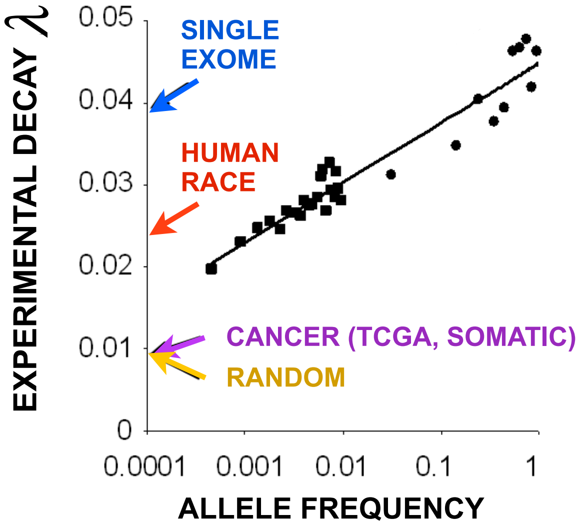
ACTION DISTRIBUTION BY ALLELE FREQUENCY



$$f'(\gamma)\cdot d\gamma = d\varphi$$

THE DECAY IS LOG-LINEAR WITH ALLELE FREQUENCY

λ=.0032ln(AF)+0.045
R^2=0.92



Fisher 1930

Katsonis and Lichtarge
Genome Research (2014)

A PERTURBATION MODEL FOR THE ACTION OF MUTATIONS ON PROTEINS

$$\boxed{f'(\gamma)\bullet d\gamma = d\phi}$$

1. Correlates with mutational impact in case studies
2. Compares well against other models
3. Exponential distribution of coding variations c/w fitness

A MULTISCALE PERTURBATION MODEL OF MUTATIONAL ACTION

$$\boxed{f'(\gamma)\bullet d\gamma = d\phi}$$

Molecular studies

- Correlates with mutational impact

Clinical studies (early)

- Correlates with disease severity in some genes

Population studies

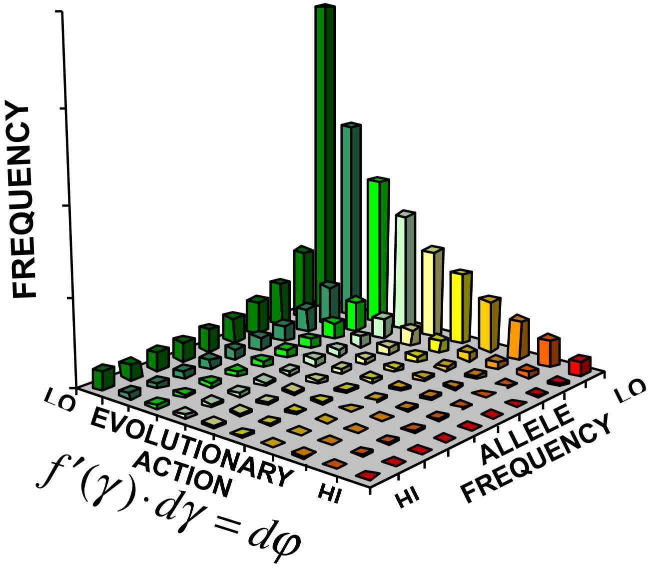
- Exponential distribution of coding variations c/w fitness
- Compares favorably to statistical/Machine Learning models

Other considerations …

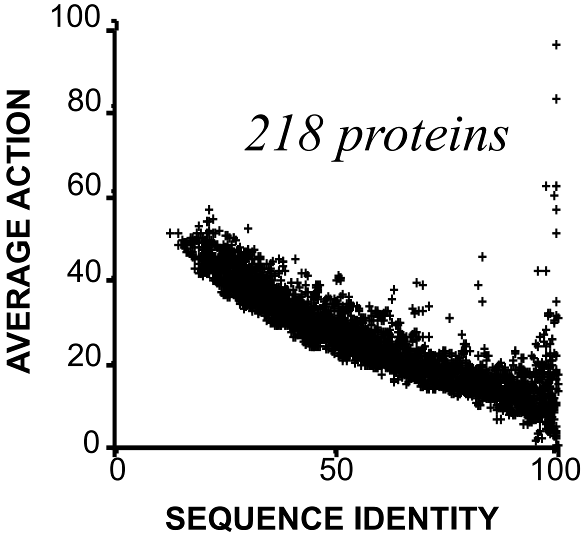
TOWARDS
A VARIATIONAL PRINCIPLE
FOR EVOLUTION?

THE ACTION OF MUTATIONS IS (NEARLY) MINIMIZED DURING EVOLUTION

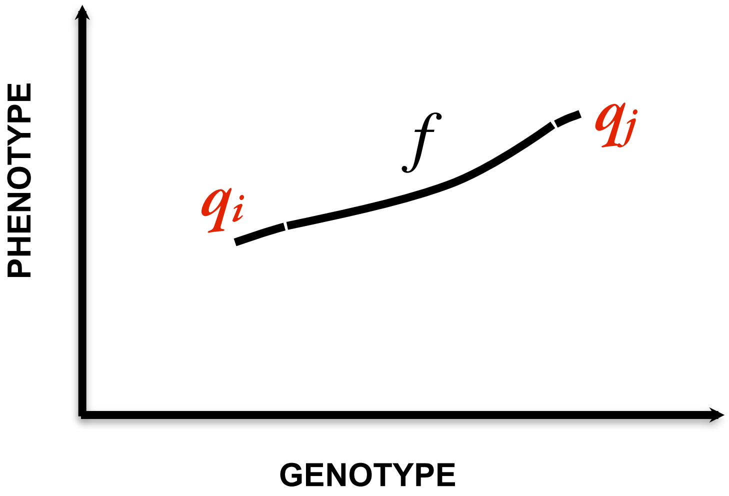
LEAST ACTION SNPs ARE THE MOST LIKELY TO CARRY ON



CLOSEST HOMOLOGS HAVE LEAST ACTION



EVOLUTIONARY PATHS SHOULD MINIMIZE ACTION AT EACH STEP



These data therefore suggest action is minimized:

$$\partial \int_{q_1}^{q_2} f'(\gamma) \bullet d\gamma = 0$$

Taking f' as the *generalized (evolutionary) momentum*

and γ as the *generalized coordinate*

leads to the usual expression of the least action principle:

$$\partial \int_{q_1}^{q_2} p \bullet dq = 0$$

Katsonis and Lichtarge
Genome Research (2014)

Summation

$$\boxed{f'(\gamma)\bullet d\gamma = d\phi}$$

CASE STUDIES AND LARGE SCALE VALIDATIONS

Application/Test	**Protein**	**Collaborator**	**References**	
Predict Binding site	Gα transducin	*Bourne*	*Lichtarge **PNAS** 1996*	*Onrust **Science** 1997*
Separate function	Nuclear Transport	*Moore*	*Cushman **J. Mol. Biol.** 2004*	
	Steroid receptors	*Smith*	*Raviscioni **Proteins** 2006*	
	Ku	*Bertuch*	*Ribes-Zamora **Nat Struc Molec Biol** 2007*	
	RecA - LexA	*Herman*	*Adikesavan **PLoS Genetics** 2011*	
Swap function	RGS	*Wensel*	*Sowa **PNAS** 2000*	*Sowa **Nat Struct Biol** 2001*
	Proneural proteins	*Hassan*	*Quan **Development** 2004*	
	Nuclear Receptors	*Cooney*	*Raviscioni **J. Mol Biol.** 2005*	
Redesign function	GPCRs	*Wensel*	*Yang **Mol Endocrinol** 2002*	*Lin **J Biol Chem**.2003*
		Lefkowitz	*Madabushi **J Biol Chem**.2004*	*Shenoy **J Biol Chem**.2006*
		Bouvier	*Kobayashi **Traffic** 2009*	
		Hansen	*Bonde **Biochem Pharmacol** 2010*	
Build inhibitors	Nuclear Receptors	*Cooney*	*Gu **J. Biol. Chem**.2005*	
	GRK	*Clark*	*Baameur **Mol Pharm** 2010*	
	Cohesin	*Pati*		
Assess gene variations	HPD	*Bodamer*	*Item **Mol Gen Metab** 2007*	
	RRM2B	*Shinawi*	*Shaibani **Arch Neurol.** 2009*	
	TP53	*Myers*	***Neskey et al J. Clin Onc.**(in revisions)*	
	CAGI			
Predict protein function	Structural Proteome		*Kristensen **Prot Sci** 2006* *Ward **PloS One** 2009* *Erdin **J Mol Biol** 2010*	*Kristensen **BMC bioinf** 2008* *Ward **Bioinformatics** 2009* *Venner **PLoS One** 2011*

A MULTISCALE PERTURBATION EQUATION OF MUTATIONAL ACTION ON FITNESS YIELDS A QUANTITATIVE MODEL BETWEEN GENOTYPE AND PHENOTYPE VARIATIONS

$$\boxed{f'(\gamma)\bullet d\gamma = d\phi}$$

Molecular studies

- Correlates with mutational impact

Clinical studies (early)

- Correlates with disease severity in some genes

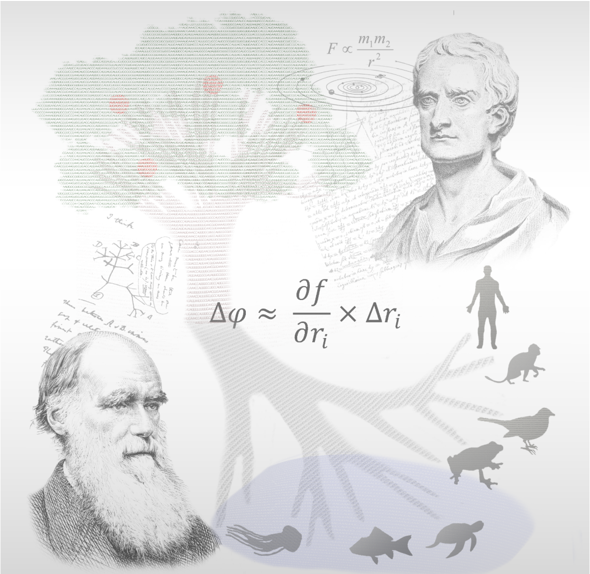
Population studies

- Exponential distribution of coding variations c/w fitness
- Compares favorably to statistical/Machine Learning models

Theory

- During evolution, the action is nearly minimized: $\partial \int_{q_1}^{q_2} f'(\gamma) \bullet d\gamma = 0$

GENOME RESEARCH



Katsonis and Lichtarge (2014)

Baylor College of Medicine
Lichtarge Lab

- Anbu Adikesavan
- Benu Atri
- Ben Bachman
- Christie Buchovecky
- Jonathan Gallion
- Teng-Kuei Hsu
- Panos Katsonis
- Amanda Koire
- Martin Lisewski
- Rhonald Lua
- David Marciano
- Ilya Novikov
- Joel Quiros
- Sam Regenbogen
- Lauren Straker
- Maria Terron
- Angela Wilkins
- Youhin Zhang
- Stephen Wilson

Ted Wensel

- Yun Min Sung
- Hye Jin Kang
- Monica Montoya

Larry Donehower

- Tajhal Dayaram
- Neha Parikh

Christophe Herman
Susan Rosenberg

- Amar Al Mamun

Huda Zoghbi

- **Steve Baker**

Rice University
Marek Kimmel

- Stephanie Hicks

MD Anderson Cancer Center
Jeff Myers

- David Neskey
- TJ Ow
- Curtis Pickering
- Abdullah Osman

Duke
Bob Lefkowitz
Marc Caron

U. De Montréal
Michel Bouvier

UTSW
Beth Levine

FUNDING

- **NIH - NIGMS**
- **NSF - DBI**
- **CPRIT**
- **GCC**
- **McNair Foundation**
- **DARPA**

Code/viewers/servers

http://mammoth.bcm.tmc.edu