CURRENT
OPINION
IN
CHEMICAL
BIOLOGY

9

2005

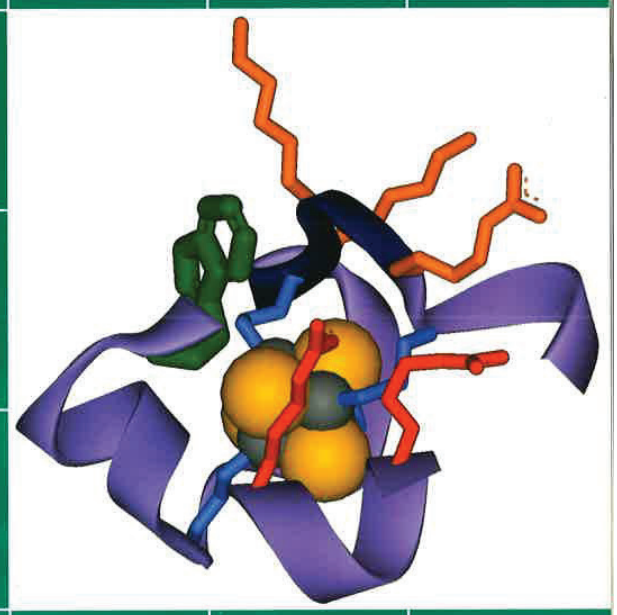
QP 550
.C87
Set 1

Vol 9 No 2 April 2005

Current Opinion in Chemical Biology

ELSEVIER

Donald Hilvert & Steven V Ley, Editors



April 2005

Bioinorganic chemistry

Edited by Elizabeth C Theil and H Holden Thorp

Biocatalysis and biotransformation

Edited by Dan E Robertson and Uwe T Bornscheuer

June 2005 Combinatorial chemistry
August 2005 Next-generation therapeutics
October 2005 Analytical techniques • Mechanisms
December 2005 Biopolymers • Model systems
February 2006 Proteomics and genomics

SCIENCE DIRECT® Access *COCH* articles online up to one month before they appear in your print journal www.sciencedirect.com

CURRENT OPINION
www.current-opinion.com

Current Opinion in **Chemical Biology**
Editors **Donald Hilvert** Switzerland **Steven V Ley** UK

Vol 9 No 2 **April 2005**

Editorial Enquiries

Current Opinion in Chemical Biology
Elsevier London
84 Theobald's Road
London WC1X 8RR, UK
Tel: +44 (0)20 7611 4400
Fax: +44 (0)20 7611 4401
e-mail: cochembiol@elsevier.com

Subscription Enquiries

e-mail: ct.subs@qss-uk.com

In-house Editor Matt Brown
Cross-title Editor Jenny Gillion
Editorial Coordinator Patrick Liu
Publishing Manager O Claire Moulton
Illustrations The Studio

Current Opinion in Chemical Biology
ISSN 1367-5931
is published bimonthly by
Elsevier Ltd
The Boulevard
Langford Lane
Kidlington
Oxford, OX5 1GB, UK

Each volume consists of six issues of approximately 130 pages.

The Aims of the journal can be found at www.elsevier.com/locate/cbi

Printed by The Manson Group Ltd, St Albans, UK.

Current Opinion in Chemical Biology
is indexed and/or abstracted by
Chemical Abstracts
Current Contents (Life Science)
Index Medicus
Life Sciences Collection
Medline/MEDLARS Online
Science Citation Index
SciSearch/Science Citation Index Expanded

Editorial Board

Duilio Arigoni (Switzerland)
Sir Jack E Baldwin (UK)
Jacqueline K Barton (USA)
Stephen J Benkovic (USA)
Ronald Breslow (USA)
David E Cane (USA)
Tom Cech (USA)
Kevin T Chapman (USA)
Anthony W Czarnik (USA)
Bill DeGrado (USA)
Anne Dell (UK)
Peter B Dervan (USA)
François N Diederich (Switzerland)
Albert Eschenmoser (Switzerland)
David A Evans (USA)
Sir Alan Fersht (UK)
Michael H Gelb (USA)
Harry B Gray (USA)
Richard Holm (USA)
Barbara Imperiali (USA)
Bryan Jones (Canada)
Gerald F Joyce (USA)
Chaitan Khosla (USA)
Judith P Klinman (USA)
Kenji Koga (Japan)
Toyoki Kunitake (Japan)
Horst Kunz (Germany)
Peter F Leadlay (UK)
Jean-Marie Lehn (France)
Stephen J Lippard (USA)
Klaus Mueller (Switzerland)
Koji Nakanishi (USA)
KC Nicolaou (USA)
Tom O'Halloran (USA)
Glenn D Prestwich (USA)
Larry Que (USA)
Julius Rebek, Jr (USA)
David N Reinhoudt (Netherlands)
Janos Rétey (Germany)
Alanna Schepartz (USA)
Stuart L Schreiber (USA)
Peter G Schultz (USA)
W Clark Still (USA)
Saul Tendler (UK)
Andrew J Thomson (UK)
Joan Selverstone Valentine (USA)
Greg Verdine (USA)
Christopher T Walsh (USA)
Chi-Huey Wong (USA)

2005 Contents

The subject of chemical biology is divided into nine major sections, each of which is reviewed once a year. Each issue contains one or two of the sections, and the amount of space devoted to each section is related to its importance.

February

Proteomics and genomics
Edited by Benjamin F Cravatt and Thomas Kodadek

April

Biocatalysis and biotransformation
Edited by Uwe Bornscheuer and Dan Robertson

Bioinorganic chemistry
Edited by Elizabeth Theil and Holden Thorp

June

Combinatorial chemistry
Edited by Prabhat Arya and Hans-Joerg Roth

August

Next-generation therapeutics
Edited by Chris Vlahos and Michael Coghlan

October

Mechanisms
Edited by Rowena Mathews and Christopher Walsh

Analytical techniques
Edited by Chris Geddes and Ramachandram Badugu

December

Biopolymers
Edited by Scott Strobel and Tom Muir

Model systems
Edited by Paolo Scrimin and Lars Baltzer

LIBRARY OF CONGRESS
APR 28 2005
E
SERIAL RECORD

DOI 10.1016/S1367-5931(05)00035-9

Current Opinion in **Chemical Biology**

Current Opinion in Chemical Biology – Online Access

Institutional Access
Full text online access to *Current Opinion in Chemical Biology* via ScienceDirect (www.sciencedirect.com) may be available as part of your institute's regular licensing arrangements.
Free abstracts to all articles are provided to non-subscribing institutes. Full text articles are available via Pay Per View.

Individual Access
Full text online access to *Current Opinion in Chemical Biology* is provided free of charge as part of an individual's or laboratory print subscription copy. Subscribers may claim their access by registering at https://cs.sciencedirect.com/activate/cochbi/members

Related information

For a full list of Elsevier products, visit www.elsevier.com

Scirus
The search engine for science, focuses only on Web pages containing scientific content. Access it now at http://www.scirus.com

Web Research Guide
Elsevier has prepared the 'Web Research Guide' to help students, faculty members, authors and research scientists find the information they need on the Web. Access it now at http://www.webresearch.sciencedirect.com/

Subscription Rates *Current Opinion in Chemical Biology*, Vol. 9, 2005, 6 issues per annum. Print prices include postage and handling.

PERSONAL	Europe (incl. UK)	North/South America[1]/ROW	Japan
Print[1]+ ONLINE ACCESS	€ 260	US$ 282	¥ 31,500
INSTITUTIONAL	Europe (incl. UK)	North/South America[1]/ROW	Japan
Print[1]	€ 1,171	US$ 1,309	¥ 162,400

[1]EU subscribers may be liable to VAT on print subscriptions. Canadian subscribers must add GST.
Subscriptions are accepted on a prepaid basis only and are entered on a calendar year basis for institutions. Issues are sent by economy (surface mail) within Europe and standard mail (air delivery) outside Europe. Claims for missing issues will be met, free of charge, if made within six months of the date of dispatch. Please send notices of change of address at least eight weeks in advance, including both old and new addresses.

Subscription orders and enquiries
All orders, renewals and changes of address should be sent directly to:
Current Opinion & Trends Subscriptions
PO Box 331, Haywards Heath, West Sussex, UK, RH16 3FG
Tel: +44 (0)1444 475650; Fax: +44 (0) 1444 445423;
e-mail: ct.subs@qss-uk.com

Orders can also be placed via a subscription agent or bookseller.

Or contact Elsevier's Regional Sales Offices:

From USA, Canada and Latin America
Contact the Orlando office: Tel: (+1) 407 345 4033; Fax: (+1) 407 365 1354
Toll Free for customers in the USA and Canada: (+1)-800-418-4348
e-mail: uscurrent@elsevier.com

From Japan
Contact the Tokyo office:
Tel: (+81) 3 5561 5033; Fax: (+81) 3 5561 5047;
e-mail: jp.info@elsevier.com

To renew your subscription, visit www.onlinerenewals.co.uk/ctprintsubs

To subscribe online, go to www.elsevier.com/locate/cbi

Commercial Sales enquiries
(including advertising in this title):
Online media pack http://mediapack.bmn.com/
e-mail: advertising@bmn.com

Reprint requests
To order reprints of this title, e-mail: currentreprints@elsevier.com

Copyright Statement

© 2005 Elsevier Ltd. All rights reserved.
This journal and the individual contributions contained in it are protected under copyright by Elsevier Ltd, and the following terms and conditions apply to their use:

Photocopying
Single photocopies of single articles may be made for personal use as allowed by national copyright laws. Permission of the publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.
Permissions may be sought directly from Elsevier's Rights Department in Oxford, UK: phone: (+44) 1865 843830; fax: (+44) 1865 853333; e-mail: permissions@elsevier.com. Requests may also be completed on-line via the Elsevier homepage (http://www.elsevier.com/locate/permissions).
In the USA, users may clear permissions and make payments through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone: (+1) (978) 7508400; fax: (+1) (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 20 7631 5555; fax: (+44) 20 7631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works
Subscribers may reproduce tables of contents or prepare lists of articles including abstracts for internal circulation within their institutions. Permission of the publisher is required for resale or distribution outside the institution.
Permission of the publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage
Permission of the publisher is required to store or use electronically any material contained in this journal, including any article or part of an article.
Except as outlined above, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the publisher.
Address permissions requests to: Elsevier Rights Department, at the fax and e-mail addresses noted above.

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.
Although all advertising material is expected to conform to ethical (medical) standards, inclusion in this publication does not constitute a guarantee or endorsement of the quality or value of such product or of the claims made of it by its manufacturer.

Current Opinion in Chemical Biology (ISSN 1367-5931) is published bi-monthly, January to December, for US$1,309, by Elsevier Ltd, The Boulevard, Langford Lane, Kidlington, UK, Oxford, OX5 1GB. Periodicals postage paid at Rahway, NJ, POSTMASTER: send address corrections to *Current Opinion in Chemical Biology* c/o BTB Mailflight Ltd, 365 Blair Rd, Avenel NJ 07001. US agent: BTB Mailflight Ltd, 365 Blair Rd, Avenel, NJ 07001.

Current Opinion in **Chemical Biology**

Vol 9 No 2 **April 2005**

Contents

Reviews

Bioinorganic chemistry
Edited by Elizabeth C Theil and H Holden Thorp

95 **Elizabeth C Theil and H Holden Thorp**
Bioinorganic chemistry

97 **Debdip Ghosh and Vincent L Pecoraro**
Probing metal–protein interactions using a *de novo* design approach

104 **Sarah A Woodson**
Metal ions and RNA folding: a highly charged topic with a dynamic future

110 **Fraser A Armstrong**
Recent developments in dynamic electrochemical studies of adsorbed enzymes and their active sites

118 **Yi Lu**
Design and engineering of metalloproteins containing unnatural amino acids or non-native metal-containing cofactors

127 **Thomas D Tullius and Jason A Greenbaum**
Mapping nucleic acid structure by hydroxyl radical cleavage

135 **Leigh JK Boerner and Jeffrey M Zaleski**
Metal complex–DNA interactions: from transcription inhibition to photoactivated cleavage

145 **Olga A Lukianova and Sheila S David**
A role for iron–sulfur clusters in DNA repair

152 **Andrea Decker and Edward I Solomon**
Dioxygen activation by copper, heme and non-heme iron enzymes: comparison of electronic structures and reactivities

Biocatalysis and biotransformation
Edited by Dan E Robertson and Uwe T Bornscheuer

164 **Dan E Robertson and Uwe T Bornscheuer**
Biocatalysis and biotransformation

166 **Haruyuki Atomi**
Recent progress towards the application of hyperthermophiles and their enzymes

174 **Takeru Ishige, Kohsuke Honda and Sakayu Shimizu**
Whole organism biocatalysis

181 **Kurt Faber and Wolfgang Kroutil**
New enzymes for biotransformations

188 **Sven Panke and Marcel Wubbolts**
Advances in biocatalytic synthesis of pharmaceutical intermediates

195 **Romas J Kazlauskas**
Enhancing catalytic promiscuity for biocatalysis

202 **Jeremy Minshull, Jon E Ness, Claes Gustafsson and Sridhar Govindarajan**
Predicting enzyme function from protein sequence

210 **Amir Aharoni, Andrew D Griffiths and Dan S Tawfik**
High-throughput screens and selections of enzyme-encoding genes

217 **Linqiu Cao**
Immobilised enzymes: science or art?

The cover

The cover figure is taken from the review on by Lukianova and David, and is based on the structure of MutY solved by Tainer and co-workers (Guan y *et al.*: **MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily.** *Nat Struct Biol* 1998, **5**:1058-1064). This figure illustrates the region around the iron–sulfur cluster of MutY that is involved in recognition of damaged DNA.

This material may be protected by Copyright law (Title 17 U.S. Code)

Full text provided by www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Predicting enzyme function from protein sequence

Jeremy Minshull, Jon E Ness, Claes Gustafsson and Sridhar Govindarajan

There are two main reasons to try to predict an enzyme's function from its sequence. The first is to identify the components and thus the functional capabilities of an organism, the second is to create enzymes with specific properties. Genomics, expression analysis, proteomics and metabonomics are largely directed towards understanding how information flows from DNA sequence to protein functions within an organism. This review focuses on information flow in the opposite direction: the applicability of what is being learned from natural enzymes to improve methods for catalyst design.

Addresses
DNA 2.0, 1455 Adams Drive, Menlo Park, CA 94025, USA

Corresponding author: Minshull, J (jminshull@dnatwopointo.com)

Current Opinion in Chemical Biology 2005, **9**:202–209

This review comes from a themed issue on
Biocatalysis and biotransformation
Edited by Dan E Robertson and Uwe T Bornscheuer

1367-5931/$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.cbpa.2005.02.003

Introduction

The sequencing of the human genome is probably more significant as a portent of the future than as the source of cures for human ills. Its three billion sequenced base pairs are a sign of the glut of information available and an incentive to interpret that information.

Hard on the heels of the genomics revolution came structural genomics and proteomics. As with the human genome sequencing project, the primary motivation behind these endeavors is to decipher the physiological functions of naturally occurring sequences. A major short-term goal for both genomics and proteomics is to identify protein drug targets and thus candidate small-molecule pharmaceuticals to either modulate the function of human proteins [1], or to selectively inhibit essential functions in pathogens [2,3].

Two new fields of study are emerging to take advantage of our accumulating understanding of how organisms are put together. The first of these is systems biology, which further extends the information-gathering philosophy and is driven by the belief that it will ultimately be possible to predict the behavior of large networks of biological elements. Complementing these research efforts that translate physical systems into information is the field of synthetic biology, which seeks to create modified or new genes, proteins and organisms, using our expanding knowledge of natural systems as a jumping-off point (Figure 1). How can synthetic biology take advantage of the information that is being gathered and processed from natural biological systems and the tools used to do so? This review aims to address ways in which studies of nature can benefit protein engineers.

Protein engineering and re-engineering

Increasing computational power is improving the ability of synthetic biologists to predict the structures into which amino acid sequences will fold [4]. Nevertheless, it is still a substantial step from predicting a protein fold to accurately designing a catalyst *de novo*. This difficulty is exemplified by studies of $(\beta/\alpha)_8$-barrel proteins: the fold is found in 21 homologous superfamilies. The common structure appears to confer some conserved properties, but there is a great diversity of function within the 76 different $(\beta/\alpha)_8$-barrel sequence families [5–7]. This difficulty for assignors of protein function is also a barrier to designers of enzymes.

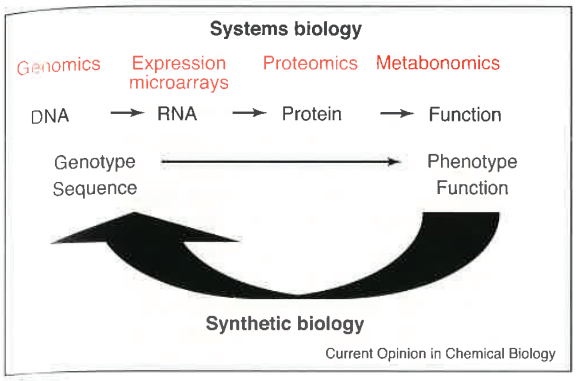
Because of the large gap that still exists between designing a protein fold and specifying a catalytic function, protein engineers have largely focused on the more modest but more tractable goal of re-engineering catalytic activity. In the most dramatic success of this approach to date, Dwyer *et al.* recently succeeded in converting a periplasmic binding protein into a triose phosphate isomerase. The authors computationally 'stripped out' the protein's ribose-binding site, and used dead-end elimination based algorithms to select amino acid mutations to produce target ligand docking and design an active site [8•]. This protein re-design required a high-resolution three-dimensional structure as a starting point. However, production of high-resolution structures still lags the rate at which sequence information is being generated. It would thus be useful if sequence information alone could guide choices of amino acid changes that might lead to desired changes in function.

Using homologous sequences for protein engineering

On microscopic timescales, the sequence of a protein determines the fold that it will adopt and thus the other molecules that it will interact with or transform. However, on macroscopic timescales it is the folding and function of proteins altered by random mutation that determines the

Corteva Exhibit 2025 - Page 6 of 13
Bayer v. Corteva - IPR2023-01038

Figure 1



Systems biology, synthetic biology and information flow. Large amounts of information from genomics, proteomics and metabonomics are analyzed to create models describing the functions of individual genes and proteins. Systems biologists are now developing methods for predicting the behavior of networks of these elements. Synthetic biology closes the loop by reversing the direction of information flow, creating novel genes, proteins and organisms based on data about naturally occurring systems, allowing tests to be made of the predictive algorithms, as well as creating new and useful compositions of matter.

elements of protein sequence that are retained (Figure 2). The sequences of orthologous proteins are thus a potentially rich source of information regarding sequence changes that are acceptable without drastically disrupting protein function, while paralogs may provide additional information about protein sequence elements responsible for functional change.

One example of the use of homologous sequence information for protein re-engineering is the approach of DNA shuffling. Homologous sequences are recombined randomly with one another, and the resultant library screened for target properties. Here, the sequence information is contained within the homologous genes. The physical PCR process used to produce recombinants ensures that recombination occurs at regions where there is substantial sequence identity [9]. Because the boundaries of recombination events are confined to DNA sequences that are reasonably conserved, the exchange sites themselves are biased against those likely to drastically disrupt protein structure or function. Nevertheless libraries of 10 000 or more recombinants must typically be screened to identify those with useful improvements.

High-throughput screening of protein function itself presents some problems, primary amongst which is the frequent lack of good correlation between a screen that is simple and one that accurately predicts function under application conditions. It would therefore be useful to develop methods for reducing the numbers of variants that must be tested.

Two methods have recently been proposed for identifying regions of proteins that may be productively transposed to a homolog. The first of these invokes the use of structural information to assess intra-protein amino acid interactions. Segments of homologous proteins are judged to be more transferable the fewer amino acid interactions are disrupted [10,11]. Although this method requires structural information, it is less demanding than the docking approach; even predicted structures can be acceptable [12•].

The second method uses only sequence information. Castle *et al.* recently created an enzyme capable of acetylating the widely used herbicide glyphosate by shuffling selected amino acids from four homologous proteins into a parental backbone. Neither the structure nor function of these donor enzymes were known. However, amino acids selected from these sequences increased the k_{cat}/K_m of the acetyl transferase from around 100 $min^{-1}mM^{-1}$ to over 8000 $min^{-1}mM^{-1}$ [13••].

Reducing the screening burden

Impressive though the glyphosate acetyl transferase example is, it required screening over 50 000 variants to achieve the necessary increase in specific activity. Preserving the successful philosophy of using substitutions found in homologs, while reducing the screening burden, requires the answers to three questions.

1. How can functional homology best be determined to identify suitable donor enzymes?
2. What is the best way to choose amino acid substitutions from these donors?
3. How can maximal information about the substitutions be derived from screening data?

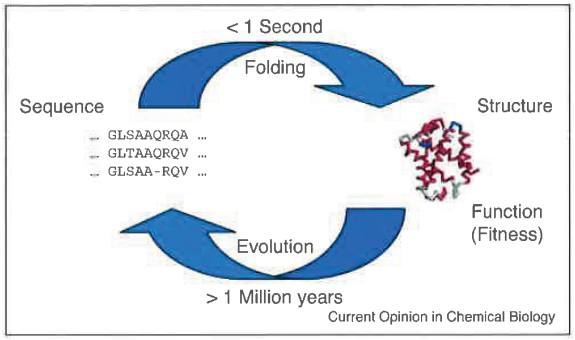
Genomics, structural biology and proteomics provide useful tools to address the first two questions. In seeking to assign function to predicted proteins, phylogenetic and structural information are frequently compared between sequences of unknown function and those whose activities have been experimentally determined. The third question may be answered by looking to traditional engineering methods, including structure–activity relationships that have found application in small-molecule drug design.

Finding homologs

Prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Despite this difficulty, the increasing availability of sequences, structures and known activities is facilitating the task. Comparative genomics permits many approaches to function prediction when the data are available. These include the use of protein-protein interaction patterns, and correlations between occurrences of related proteins in different organisms

Corteva Exhibit 2025 - Page 7 of 13
Bayer v. Corteva - IPR2023-01038

Figure 2



Sequence determines function, function constrains sequence. The sequence of a protein's amino acids, together with environmental factors such as other proteins within a cell determines its fold, typically in less than a second. The three-dimensional structure of the protein in turn governs its function. Random changes to the protein sequence occur over evolutionary timescales; changes are retained if they are functionally equivalent or confer a selective advantage.

[14–17]. For example it is now known that members of mechanistically diverse superfamilies share the ability to catalyze a common partial reaction [18]. Single amino acid changes near the active site of L-Ala-D/L-Glu epimerase and muconate lactonizing enzyme converted these proteins into catalysts for a third reaction catalyzed by another superfamily member, *o*-succinyl benzoate synthase [19••]. A better understanding of the catalytic promiscuity of natural enzymes would undoubtedly aid in protein re-engineering [20].

The conservation of structure and partial function between superfamily members makes them appear attractive substitution donors and acceptors. A new structure–function linkage database (http://sfld.rbvi.ucsf.edu) aims to store sequence–structure–function information that can be interrogated with desired chemical reactions [21]. This will likely be a valuable resource once it is sufficiently populated.

Analysis of residue conservation and variation among homologous sequences that have evolved under similar constraints also provides information about residues that confer desired properties. For example, in one recent study, the consensus amino acid sequence of 13 fungal phytases was synthesized and found to be thermostable at temperatures between 13 and 26 °C higher than any of the 13 wild-type phytases used to create the consensus. Addition of more phytase sequences resulted in a new consensus sequence with further improved thermostability [22•].

Superfamily members often share relatively low sequence identities. In the absence of reliable functional assignments, probing public databases for proteins with reasonable amino acid identity is the simplest method for identifying potential donors. When the functional similarity of 486 000 homologous protein pairs were compared, very little variation in EC number was observed in sequences sharing more than 40% identity [23]. In the glyphosate acetyl transferase experiments, enzymes were selected purely on the basis of sequence identity with the acceptor gene [13••]. The amino acid sequence identity of these donors ranged from 49% to as low as 28%, the two most homologous of these donors catalyzed the transferase reaction less well than the acceptor, the two least homologous were not even tested for function. Despite this apparent functional unsuitability, the two more homologous donors contributed 13 amino acid substitutions and the two least homologous 5 substitutions to an 800-fold improvement of k_{cat}/K_m.

Choosing substitutions

Many computational methods are now being tested for their ability to identify functionally important residues. These residues can then be used to help functional assignment of proteins, or to predict regions or residues involved in protein–protein interactions or catalysis. Functional residue predicting methods typically use one or more of the following strategies:

1. Use of phylogenetic analysis to define evolutionary distances between homologs [24••] and cluster them based on evolutionarily related sub-families. Evolutionary trace, sequence motif and conserved functional group analysis of homologous proteins to identify putative active sites [25–28].
2. Degree of amino acid conservation at positions in related proteins [24••,29•].
3. Use of principal component analysis of sequences [24••,30].
4. Use of structural information such as proximity of observed changes to active site [29•].
5. Use of structural information such as conservation of specific residues in specific three-dimensional positions [31•,32,33].
6. Use of substitution matrices that describe the relative likelihood of amino acid conversions [31•,34,35].
7. Analysis of rates of synonymous and non-synonymous substitutions in the genes and identifying nature of selection (adaptive or neutral) undergone by the gene or specific sites in the protein [36,37].
8. Identification of functional domains using sequence- and structure-based methods [16,38].

In general, these methods are tuned to recognize amino acids whose alteration would be deleterious to enzyme function. This information can also assist in achieving the balance required to alter amino acids that will affect protein function but not destroy it, important in designing a new enzyme. For example, changes at key active site

Corteva Exhibit 2025 - Page 8 of 13
Bayer v. Corteva - IPR2023-01038

residues or other highly conserved residues are generally deleterious [29•]. Changes that are found in highly related proteins are likely to be acceptable, even if close to an active site [29•]. The effect of amino acid substitutions on protein function can also be inferred from substitution matrices derived from natural proteins. Substitution matrices are formal representations of the probability that one amino acid may be replaced by another within a functional protein. Such tables can be derived from sources ranging from very large sets of proteins [39,40] to smaller and functionally related families [41], to specific regions of proteins [35,42], including those that are most conserved [34]. They can also be extended to incorporate the frequency with which dipeptides occur [43], or be based on the physicochemical properties of residues [44].

Evolutionary conservation seems to have been a consideration in the choice of substitutions to use in the shuffling of glyphosate acetyl transferase [13••]. The authors appear to have made substitutions in the target gene that were present in one of the homologs (28–49% amino acid identity). The majority of these substitutions are also those that substitution matrices indicate have a high generic likelihood of being able to functionally replace one another, as shown in Table 1. For example, in a large number of closely related proteins, isoleucine is often replaced by leucine and vice versa. Likewise, other amino acid pairs that share chemical similarities are often exchanged in highly homologous proteins: aspartate and glutamate are both acidic, arginine and lysine both basic, serine and threonine both have hydroxyl groups and tyrosine and phenylalanine are both aromatic. By constraining substitution choice in this way, the number of candidate substitutions is drastically reduced: only a dozen or so of the possible 380 changes (19 possible changes for each of 20 amino acids) found in homologs are considered. This constraint increases the likelihood of selecting changes that would have been found in closer homologs than those actually available. Substitutions that are evolutionarily closer are more likely to retain structure and function when placed in the context of a homolog [45]. It is also clear from the glyphosate example that conservative substitutions still provide sufficient variation to achieve very considerable increases in activity (see column 3 in Table 1). From this success, it seems that the types of combined calculation methods to identify regions of conservation that are being developed for functional assignment of proteins [24••,31•] will also be useful for the choice of substitutions for protein engineering projects.

Table 1

Evolutionary conservation as a guide for functional compatibility.

Interconverted pair	Occurrences in library	Occurrences in 'winner'	Dayhoff PAM 100 matrix scores (rank/190)
I and L	9	5	2.61 (8)
D and E	6	1	3.37 (4)
R and K	7	4	3.54 (3)
S and T	4	2	2.60 (9)
Y and F	6	3	5.25 (1)
V and I	6	4	4.21 (2)
L and M	5	1	3.30 (5)
E and Q	2	1	2.46 (10)
D and N	1	0	2.41 (12)
M and I	1	0	2.76 (7)
M and V	1	1	0.64 (24)
Q and K	0	1	1.92 (14)
L and T	0	1	−4.02 (110)
G and S	0	1	−0.99 (41)
R and G	0	2	−3.55 (99)
V and A	0	1	−0.23 (28)
T and I	0	1	−2.16 (120)

Thirty independent changes were incorporated into the 146 amino acid glyphosate acetyl transferase protein, resulting in a 10 000-fold improvement in k_{cat}/K_m. The table shows the frequency with which specific amino acid interconversions were represented in the initial libraries and in a sequenced highly active shufflant. A change of I to L and a change of L to I are both included as interconverted pair I and L.

Predicting function from sequence

One of the most exciting advances in elucidating the relationship between sequence and function is the use of dimension-compressing techniques such as principal component analysis to identify key sequence elements that correlate with specific activities. These techniques were pioneered to assist in functional assignments for naturally occurring proteins [30]. However, they are also capable of identifying the specific amino acids that appear to be involved in those functions [46]. In this way they are similar to the combined phylogenetic and structural approaches mentioned above [24••,31•].

Unlike combined phylogeny/structural methods, however, dimension-compressing sequence methods also allow construction of quantitative relationships between sequence and activity. This is extremely useful for sequence variants that have been synthesized and tested in the laboratory. Sequence–activity relationships may be derived in three steps. First, amino acid sequences are encoded, for example as a matrix as shown in Table 2 and Table 3. Second, an equation is produced for each sequence, an example is the form shown in Equation 1. Third, values of c_i are chosen such that the experimentally measured values of y are as close as possible to those predicted using Equation 1.

$$y = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + \ldots c_ix_i \quad (1)$$

y: the experimentally measured activity

x_i: the amino acid variation at position i, as shown in Table 2 and 3

c_i: a weighting factor

Corteva Exhibit 2025 - Page 9 of 13
Bayer v. Corteva - IPR2023-01038

Table 2

Encoding sequences for sequence-activity determination.

Sequence #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	A	G	R	W	G	I	G	A	Y	H	K	L	I	M	A
2	A	G	R	T	G	V	G	V	Y	H	K	L	I	M	A
3	A	G	R	W	G	V	G	V	Y	H	R	L	I	M	A
4	A	G	R	T	G	M	G	A	Y	H	R	L	I	M	A

Four 15 amino acid sequences contain differences at 4 positions. (See also Table 3.)

The most widely used method for minimizing differences between experimentation and prediction is partial least squares projection onto latent variables, although there are many other possible methods. Once the best values for c_i have been calculated, validation of the relationship can be visualized graphically (Figure 3). The substitutions with the highest weighting factors are those that are predicted to be the most important for the measured function. The simplest optimization method is thus to combine all of the substitutions that are determined to be favorable [47••].

Equation 1 is linear in form; it does not take into account possible interactions between amino acids. In reality, it is reasonable to expect that an amino acid change at one position may affect the amino acids that are acceptable at one or more other positions. By selecting changes that are found in homologs, the chance of choosing a change that is incompatible with overall protein structure and function is greatly reduced. Even changes that appear to occur together, that is, both are found in some homologs and neither appear in others, may reflect the relative paucity of sequence information rather than a true functional constraint. In a recent study 52 amino acid changes found within a 180 amino acid region of 15 naturally occurring subtilisin variants were randomly recombined. Of the 100 amino acid pairs that appeared to co-vary in the parental sequences, only 7 of these appeared to result from functional constraints [48•]. In another study, of 14 substitutions that improved thermostability of prolyl endopeptidase, in only one amino acid pair (out of a total of 91 pairs) did the substitution at one position affect the choice of substitution at one other position [49•]. Such a low level of functional interaction allows relatively simple modifications of Equation 1, for example the addition of a single interactive term, to accurately model sequence–activity relationships while still using only small datasets. With or without interactive terms, sequence–activity models have been used successfully to identify and combine functionally beneficial residues to optimize peptides and proteins [47••,49•,50–52].

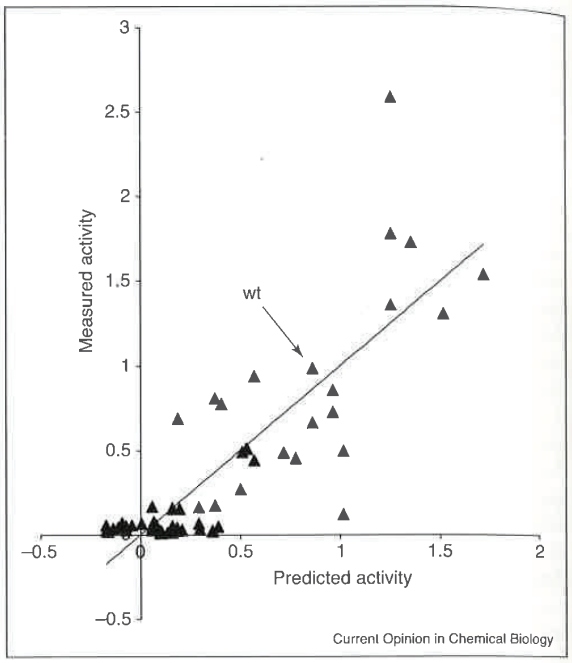
Table 3

Encoding sequences for sequence–activity determination, continued.

	4T	4W	6I	6V	6M	8A	8V	11K	11R
Sequence #	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0	1	1	0	0	1	0	1	0
2	1	0	0	1	0	0	1	1	0
3	0	1	0	1	0	0	1	0	1
4	1	0	0	0	1	1	0	0	1

These differences can be encoded as a matrix, where each possible amino acid at each variable position occupies a column. In more complex encoding strategies, columns may represent physicochemical properties of the amino acids [54].

Figure 3



Validating a sequence–activity relationship. Predicted values for a set of sequences (*x*-axis) are plotted against those experimentally determined (*y*-axis). The data are taken from [47••]. The better the correlation, the more closely the data lies on a straight line. Activity is expressed as relative to wild type (wt).

The small number of variants that must be tested to derive a predictive sequence–activity model has the tremendous advantage that these variants can be tested under conditions that are much more faithful to the final application than is generally possible in a high-throughput screen.

There are currently few published examples in which sequence–activity relationships have been used for protein optimization. The major reason for this is probably that the approach has been limited by molecular biology: it has been far easier to synthesize libraries, for example by mutagenic PCR or DNA shuffling, than it has been to

Corteva Exhibit 2025 - Page 10 of 13
Bayer v. Corteva - IPR2023-01038

create even a modest set of 50 designed genes. Indeed, it has been proposed that sequence–activity relationships could be used to analyze active variants in such libraries, enabling much smaller focused libraries to be synthesized [53•]. One strong caveat to such an approach is the frequent lack of correlation between activities measured by high-throughput screens and those actually required in an application.

In contrast to the library-screening approach, Ness *et al.* recently constructed a partial least squares regression model correlating the sequences of 48 individually designed and synthesized proteinase K variants with three different but related proteolytic activities. One of these sequence–activity models was experimentally validated: a variant that combined all of the amino acid changes that the model identified as beneficial was found to be more active than the test set, as predicted by the model [47••]. Models for the other two proteolytic activities predicted quite different beneficial amino acid changes. Consistent with the experiences of many high-throughput screeners, this suggests that extreme caution is needed when optimizing protein performance using a surrogate screen. Since sequence–activity relationships require the testing of orders of magnitude fewer variants than traditional empirical approaches [47••,53•], the use of such predictive algorithms will permit researchers to design lower throughput screens that more accurately reflect the properties they are seeking.

Conclusions

Most tools for predicting enzyme function from amino acid sequence have been developed as part of the ongoing effort to decode the genomes of humans and pathogens for health-related goals. With relatively small adaptations, these tools may now be used to aid in enzyme redesign: identifying suitable proteins as catalytic starting points, choosing changes to make, and assessing the effects of those changes in a powerfully predictive way. In particular, the use of sequence–activity relationships should free protein engineers from the expenses and inaccuracies of high-throughput screens, and allow them to directly measure the protein properties that they really care about.

Acknowledgements

We thank Patsy Babbitt for discussions regarding the use of superfamily information for modifying protein function.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Onyango P: **The role of emerging genomics and proteomics technologies in cancer drug target discovery**. *Curr Cancer Drug Targets* 2004, **4**:111-124.

2. Wang Y, Wu Y: **Computer assisted searches for drug targets with emphasis on malarial proteases and their inhibitors.** *Curr Drug Targets Infect Disord* 2004, **4**:25-40.

3. Wang S, Eisenberg D: **Crystal structures of a pantothenate synthetase from M. tuberculosis and its complexes with substrates and a reaction intermediate.** *Protein Sci* 2003, **12**:1097-1108.

4. Russ W, Ranganathan R: **Knowledge-based potential functions in protein design.** *Curr Opin Struc Biol* **12**:447-452.

5. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.** *J Mol Biol* 2002, **321**:741-765.

6. Hocker B, Claren J, Sterner R: **Mimicking enzyme evolution by generating new $(\beta\alpha)_8$-barrels from $(\beta\alpha)_4$-half-barrels.** *Proc Natl Acad Sci USA* 2004, **101**:16448-16453.

7. Sterner R, Schmid FX: **Biochemistry. *De novo* design of an enzyme.** *Science* 2004, **304**:1916-1917.

8. • Dwyer MA, Looger LL, Hellinga HW: **Computational design of a biologically active enzyme.** *Science* 2004, **304**:1967-1971.
Computational redesign of a ribose-binding protein using dead-end elimination algorithms. Around 20 mutations are introduced to convert the binding protein into a highly active triose phosphate isomerase enzyme.

9. Moore GL, Maranas CD, Lutz S, Benkovic SJ: **Predicting crossover generation in DNA shuffling.** *Proc Natl Acad Sci USA* 2001, **98**:3226-3231.

10. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG, Arnold FH: **Library analysis of SCHEMA-guided protein recombination.** *Protein Sci* 2003, **12**:1686-1693.

11. Otey CR, Silberg JJ, Voigt CA, Endelman JB, Bandara G, Arnold FH: **Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach.** *Chem Biol* 2004, **11**:309-318.

12. • Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH: **Protein building blocks preserved by recombination.** *Nat Struct Biol* 2002, **9**:553-558.
A useful method for designing chimeric proteins that are too distant to allow DNA shuffling. Recombination points are chosen based on minimizing the number of intramolecular contacts disrupted.

13. •• Castle LA, Siehl DL, Gorton R, Patten PA, Chen YH, Bertain S, Cho HJ, Duck N, Wong J, Liu D *et al.*: **Discovery and directed evolution of a glyphosate tolerance gene.** *Science* 2004, **304**:1151-1154.
A very instructive case study for the use of information found in homologous sequences to achieve dramatic increases in enzyme activity. By incorporating selected changes from distant homologs, the authors increase k_{cat}/K_m for glyphosate acetylation 10 000-fold. The online supplementary material is particularly informative.

14. Whisstock JC, Lesk AM: **Prediction of protein function from protein sequence and structure.** *Q Rev Biophys* 2003, **36**:307-340.

15. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1996, **96**:2896-2901.

16. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L: **Functional inferences from reconstructed evolutionary biology involving rectified databases–an evolutionarily grounded approach to functional genomics.** *Res Microbiol* 2000, **151**:97-106.

17. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

18. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70**:209-246.

19. •• Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, Babbitt PC, Minshull J, Gerlt JA: **Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily.** *Biochemistry* 2003, **42**:8387-8393.
Conversion of muconate lactonizing enzyme and L-Ala-D/L-Glu epimerase into *o*-succinyl benzoate synthases by introduction of single substitutions

Corteva Exhibit 2025 - Page 11 of 13
Bayer v. Corteva - IPR2023-01038

demonstrates the power of the superfamily approach to enzyme re-engineering.

20. Bornscheuer UT, Kazlauskas RJ: **Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways**. *Angew Chem Int Ed Engl* 2004, **42**:6032-6040.

21. Pegg SC-H, Brown SA, Ojha S, Huang CC, Ferrin TE, Babbitt PC: **Representing structure-function relationships in mechanistically diverse enzyme superfamilies**. *Pacific Symposium on Biocomputing* 2005, in press.

22. • Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept**. *Protein Eng* 2002, **15**:403-411.

Homologous phytase sequence information is used to build a consensus sequence with increased thermostability. Analysis of individual changes shows that the improvement results from a combination of many small increases in thermostability from each amino acid substitution.

23. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**:1113-1143.

24. •• del Sol Mesa A, Pazos F, Valencia A: **Automatic methods for predicting functionally important residues**. *J Mol Biol* 2003, **326**:1289-1302.

Excellent description of methods for using phylogenetic trees, substitution matrices and principal component analysis to identify functionally important amino acids.

25. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O: **An accurate, sensitive, and scalable method to identify functional sites in protein structures**. *J Mol Biol* 2003, **326**:255-261.

26. Lichtarge O, Yao H, Kristensen DM, Madabushi S, Mihalek I: **Accurate and scalable identification of functional sites by evolutionary tracing**. *J Struct Funct Genomics* 2003, **4**:159-166.

27. Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O: **Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions**. *J Biol Chem* 2004, **279**:8126-8132. Epub 2003 Dec 8105.

28. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O: **Structural clusters of evolutionary trace residues are statistically significant and common in proteins**. *J Mol Biol* 2002, **316**:139-154.

29. • Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, Speir JA, Fetrow JS, Baxter SM: **Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors**. *Proteins* 2003, **53**:806-816.

Combination of structural, phylogenetic and substitution matrix information to predict which substitutions are likely to be deleterious.

30. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins**. *Nat Struct Biol* 1995, **2**:171-178.

31. • Chelliah V, Chen L, Blundell TL, Lovell SC: **Distinguishing structural and functional restraints in evolution in order to identify interaction sites**. *J Mol Biol* 2004, **342**:1487-1504.

Identification of functional constraints within protein sequences using analysis of conserved residues, substitution frequencies and three-dimensional structure.

32. Innis CA, Anand AP, Sowdhamini R: **Prediction of functional sites in proteins using conserved functional group analysis**. *J Mol Biol* 2004, **337**:1053-1068.

33. Innis CA, Shi J, Blundell TL: **Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis**. *Protein Eng* 2000, **13**:839-847.

34. Friedberg I, Margalit H: **Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function**. *Protein Sci* 2002, **11**:350-360.

35. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL: **Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds**. *Protein Sci* 1992, **1**:216-226.

36. Ina Y: **New methods for estimating the numbers of synonymous and nonsynonymous substitutions**. *J Mol Evol* 1995, **40**:190-226.

37. Suzuki Y, Gojobori T, Nei M: **ADAPTSITE: detecting natural selection at single amino acid sites**. *Bioinformatics* 2001, **17**:660-661.

38. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains**. *Brief Bioinform* 2002, **3**:246-251.

39. Benner SA, Cohen MA, Gonnet GH: **Amino acid substitution during functionally constrained divergent evolution of protein sequences**. *Protein Eng* 1994, **7**:1323-1332.

40. Dayhoff MO, Eck FV: **A model of evolutionary change in proteins**. *Atlas of Protein Sequence and Structure* 1968, **3**:33-41.

41. Dimmic MW, Rest JS, Mindell DP, Goldstein RA: **rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny**. *J Mol Evol* 2002, **55**:65-73.

42. Koshi JM, Goldstein RA: **Context-dependent optimal substitution matrices**. *Protein Eng* 1995, **8**:641-645.

43. Gonnet GH, Cohen MA, Benner SA: **Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix**. *Biochem Biophys Res Commun* 1994, **199**:489-496.

44. Koshi JM, Mindell DP, Goldstein RA: **Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes**. *Mol Biol Evol* 1999, **16**:173-179.

45. Kurtzman AL, Govindarajan S, Vahle K, Jones JT, Heinrich V, Pattent PA: **Advances in directed protein evolution by recursive genetic recombination: applications to therapeutic proteins**. *Curr Opin Biotechnol* 2001, **12**:361-370.

46. Gogos A, Jantz D, Senturker S, Richardson D, Dizdaroglu M, Clarke ND: **Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: an experimental test using DNA glycosylase homologs**. *Proteins* 2000, **40**:98-105.

47. •• Ness JE, Cox AJ, Govindarajan S, Gustafsson C, Gross RA, Minshull J: **Empirical biocatalyst engineering: escaping the tyranny of high throughput screening**. In *ACS Symposium Series 900 Polymer Biocatalysis and Biomaterials*. Edited by Cheng HA, Gross RA. American Chemical Society; 2005, in press, vol. 900.

Amino acid substitutions were chosen using multiple parameters including phylogenetic information and substitution matrices. Each change is assessed for effect on three functional properties using sequence–activity relationships, and an improved variant is synthesized from this information.

48. • Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C: **Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation**. *J Mol Biol* 2003, **328**:1061-1069.

Active subtilisin variants in which all parental covariation is removed are found to contain almost no co-varying amino acid pairs.

49. • Aita T, Hamamatsu N, Nomiya Y, Uchiyama H, Shibanaka Y, Husimi Y: **Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution**. *Biopolymers* 2002, **64**:95-105.

Sequence–activity modeling of prolyl endopeptidase is used to identify a single pair of co-dependent amino acid substitutions out of 91 possible pairs.

50. Prusis P, Lundstedt T, Wikberg JE: **Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors**. *Protein Eng* 2002, **15**:305-311.

51. Norinder U, Rivera C, Unden A: **A quantitative structure-activity relationship study of some substance P-related peptides. A multivariate approach using PLS and variable selection**. *J Pept Res* 1997, **49**:155-162.

52. Bucht G, Wikstrom P, Hjalmarsson K: **Optimising the signal peptide for glycosyl phosphatidylinositol modification of**

Corteva Exhibit 2025 - Page 12 of 13
Bayer v. Corteva - IPR2023-01038

human acetylcholinesterase using mutational analysis and peptide-quantitative structure-activity relationships. *Biochim Biophys Acta* 1999, **1431**:471-482.

53. • Fox R, Roy A, Govindarajan S, Minshull J, Gustafsson C, Jones JT, Emig R: **Optimizing the search algorithm for protein engineering by directed evolution.** *Protein Eng* 2003, **16**:589-597.

A nice theoretical treatment of protein fitness landscapes, showing the dramatic decrease in screening that can be achieved using sequence–activity relationships.

54. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids.** *J Med Chem* 1998, **41**:2481-2491.

Corteva Exhibit 2025 - Page 13 of 13
Bayer v. Corteva - IPR2023-01038