Ex. 1044 Nicola 1977

A 20703 D

FEBS

Volume 78 Number 1 August (II) 1977

LINDA HALL
SEP 13 1977
LIBRARY

European Journal of Biochemistry

Editorial Board

Honorary Chairman: Sir Hans Krebs (Oxford)

Editor-in-Chief: Claude Liébecq (Liège)

Associate Editors: Giorgio Bernardi (Paris), J. Ieuan Harris (Cambridge), Lothar Jaenicke (Köln), Cees Veeger (Wageningen)

Editors: F. Franěk (Praha), G. P. Georgiev (Moskva), H. G. Hers (Bruxelles), T. Keleti (Budapest), M. Lazdunski (Nice), P. Overath (Tübingen), J. G. Reich (Berlin), M. Revel (Rehovot), M. Sela (Rehovot), C. A. Vernon (London), J.-P. Waller (Palaiseau), D. H. Williamson (Oxford), H. G. Zachau (München), W. Zillig (Martinsried)

Special Advisor to the Editor-in-Chief: H. Mayer-Kaupp (Heidelberg)

Advisory Board

C. Baglioni (Albany), R. Banerjee (Paris), E. K. F. Bautz (Heidelberg), H. Beaufay (Bruxelles), L. D. Bergelson (Moskva), L. Boross (Szeged), F. Bossa (Camerino), E. M. Bradbury (Portsmouth), R. Braun (Bern), H. Buc (Paris), P. H. W. Butterworth (London), E. Carafoli (Zürich), M. J. Clemens (London), Sir John Cornforth (Brighton), K. Decker (Freiburg i. Br.), G. H. de Haas (Utrecht), W. Doerfler (Köln), J.-P. Ebel (Strasbourg), W. C. Fiers (Gent), C. Gancedo (Madrid), P. B. Garland (Dundee), J.-M. Ghuysen (Liège), D. B. Gower (London), D. O. Hall (London), E. J. M. Helmreich (Würzburg), P. Hemmerich (Konstanz), B. Hirt (Lausanne), E. Hofmann (Leipzig), O. Hoffmann-Ostenhof (Wien), R. Jaenicke (Regensburg), P. Jollès (Paris), M. Ya. Karpeisky (Moskva), E. Katchalski-Katzir (Rehovot), K. Kirschner (Basel), M. Klingenberg (München), H. Klostermeyer (Kiel), G. L. E. Koch (Cambridge), L. Kovac (Ivánka pri Dunaji), C. G. Kurland (Uppsala), U. G. Lagerkvist (Göteborg), A. Levitzki (Jerusalem), U. Z. Littauer (Rehovot), R. Lontie (Leuven), O. Lüderitz (Freiburg i. Br.), S. Magnusson (Aarhus), C. Milstein (Cambridge), R. Monier (Villejuif), K. Mosbach (Lund), E. A. Newsholme (Oxford), D. Oesterhelt (Würzburg), S. Orrenius (Stockholm), T. A. J. Payens (Ede), D. W. G. Pette (Konstanz), R. J. Planta (Amsterdam), J. Ricard (Marseille), R. Rigler (Stockholm), A. Schellenberger (Halle/Saale), M. C. Scrutton (London), G. Semenza (Zürich), A. Sentenac (Saclay), S. Shaltiel (Rehovot), N. Sharon (Rehovot), D. Shugar (Warszawa), F. Šorm (Praha), F. S. Steven (Manchester), H. Sund (Konstanz), R. Thauer (Marburg/Lahn), R. Thomas (Rhode-St-Genèse), K. F. Tipton (Cambridge), A. Trebst (Bochum), D. R. Trentham (Philadelphia), R. Tsanev (Sofiya), V. Ullrich (Homburg), K. van Dam (Amsterdam), D. Vazquez (Madrid), L.-V. von Stedingk (Stockholm), V. P. Whittaker (Göttingen), O. Wieland (München), M. Wiewiorowski (Poznan), L. Wojtczak (Warszawa)

Published by Springer-Verlag Berlin Heidelberg New York on behalf of the Federation of European Biochemical Societies

Eur. J. Biochem.
ISSN 0014-2956 EJBCAI 78(1) 1–316 (1977)

August 15, 1977
published semimonthly

European Journal of Biochemistry

Volume 78 Number 1 August (II) 1977

Contents

1 Repetitive and Non-repetitive Sequences in the Transcript *in vitro* of Porcine Thyroid Chromatin
V. E. Avvedimento, A. M. Acquaviva, and S. Varrone

11 Hydrophobic and Carbohydrate-Recognition Chromatographies of Collagen Glucosyltransferase
H. Anttinen, R. Myllylä, and K. I. Kivirikko

19 Purification and Characterization of Human Antiplasmin, the Fast-Acting Plasmin Inhibitor in Plasma
B. Wiman and D. Collen

27 The Isolation of Tubulin and Actin from Mouse 3T3 Cells Transformed by Simian Virus 40 (SV3T3 Cells), an Established Cell Line Growing in Culture
K. Weber, R. Koch, W. Herzog, and J. Vandekerckhove

33 Glycolipids of the Human Gastric Content. Structure of the Sulfated Glyceroglucolipid
B. L. Slomiany, A. Slomiany, and G. B. J. Glass

41 Carbohydrate Incorporation in Plasma Membranes of Mouse Thymocytes Stimulated by Concanavalin A
R. V. W. van Eijk and P. F. Mühlradt

55 Specificity of Elongation Factor Tu from *Escherichia coli* with Respect to Attachment of the Amino Acid to the 2′ or 3′-Hydroxyl Group of the Terminal Adenosine of tRNA
M. Sprinzl, M. Kucharzewski, J. B. Hobbs, and F. Cramer

63 Refinement of the Solution Conformation of Valinomycin with the Aid of Coupling Constants from the ^{13}C-Nuclear-Magnetic-Resonance Spectra
V. F. Bystrov, Yu. D. Gavrilov, V. T. Ivanov, and Yu. A. Ovchinnikov

83 Induction of Division Synchrony in *Tetrahymena pyriformis* by a Single Hypoxic Shock. Its Use in Elucidating Control of the Cell Cycle by Adenosine 3′:5′-Monophosphate
J. R. Dickinson, M. G. Graves, and B. E. P. Swoboda

89 The Incorporation of [^{14}C]Glucosamine into Dolichol Diphosphate *N*-Acetyl[^{14}C]glucosamine by Unbroken Liver Cells in Culture
J. R. Cooper and F. W. Hemming

95 Studies on Nucleotidases in Plants. Isolation and Properties of the Monomeric Form of the Crystalline and Homogeneous Mung Bean Nucleotide Pyrophosphatase
C. V. Balakrishnan, C. S. Vaidyanathan, and N. A. Rao

103 Chorismate Mutase/Prephenate Dehydratase from *Escherichia coli* K12. Modification with 5,5′-Dithio-bis(2-nitrobenzoic acid)
M.-J. H. Gething and B. E. Davidson

111 Chorismate Mutase/Prephenate Dehydratase from *Escherichia coli* K12. Effects of Chemical Modification on the Enzymic Activities and Allosteric Inhibition
M.-J. H. Gething and B. E. Davidson

119 Purification, Crystallisation and Preliminary X-Ray Studies on Avian Pancreatic Polypeptide
S. P. Wood, J. E. Pitts, T. L. Blundell, I. J. Tickle, and J. A. Jenkins

127 The Interactive Binding of Two Ligands by an Allosteric Protein
R. F. Steiner and L. Greer

133 Structural Rearrangements Due to Ligand Binding and Haem Replacement in Myoglobin and Leghaemoglobins
N. A. Nicola and S. J. Leach

141 Isolation and Characterization of Two Methionine : tRNA Ligases from Wheat Germ
M. D. Rosa and P. B. Sigler

153 Physical Studies on the Conformation of Ribosomal Protein S4 from *Escherichia coli*
C. A. Morrison, R. A. Garrett, and E. M. Bradbury

161 Demonstration of Two α-Globin Genes per Human Haploid Genome for Normals and Hb J Mexico
P. Tolstoshev, R. Williamson, J. Eskdale, G. Verdier, J. Godet, V. Nigon, G. Trabuchet, and M. Benabadji

(Continuation overleaf)

Contents (Continuation)

167 Microtubule Assembly *in vitro*. Purification of Assembly-Promoting Factors
A. Fellous, J. Francon, A.-M. Lennon, and J. Nunez

175 Reverse Transcription of Thyroglobulin 33-S mRNA
C. Dinsart, F. van Voorthuizen, and G. Vassart

183 Influence of 1,2,3-Benzene-tricarboxylate on Pyruvate Metabolism in Rat-Liver Mitochondria
J. W. Stucki

189 Phytohaemagglutinin Stimulation of Human Lymphocytes during Amino-Acid Deprivation. RNA Polymerase I Activity of Isolated Nuclei
C. Dauphinais and W. I. Waithe

195 The Size and Shape of Human and Bovine Antithrombin III
B. Nordenman, C. Nyström, and I. Björk

205 Electron Microscopy Analysis of the Interaction between *Escherichia coli* DNA-Dependent RNA Polymerase and the Replicative Form of Phage fd DNA. 1. Mapping of the Binding Sites
P. U. Giacomoni, E. Delain, and J. B. Le Pecq

215 Electron Microscopy Analysis of the Interaction between *Escherichia coli* DNA-Dependent RNA Polymerase and the Replicative Form of Phage fd DNA. 2. Analysis of the Dissociation Kinetics
P. U. Giacomoni, E. Delain, and J. B. Le Pecq

221 Prostatic Binding Protein. A Steroid-Binding Protein Secreted by Rat Prostate
W. Heyns and P. de Moor

231 Studies on the Stimulation by Ca^{2+} and the Inhibition by ADP of Steroid 11β-Hydroxylation in Adrenal Mitochondria
A. M. Moustafa and S. B. Koritz

239 Modification of L-Isoleucyl-tRNA Synthetase with L-Isoleucyl-Bromomethyl Ketone. The Effect on the Catalytic Steps
P. Rainey, B. Hammer-Raber, M.-R. Kula, and E. Holler

251 Studies on Protein-Phosphorylation Reactions in Isolated Chromatin
J. Böhm, G. Keil, and R. Knippers

267 Functional Roles of 50-S Ribosomal Proteins
F. Hernández, D. Vázquez, and J. P. G. Ballesta

273 Purification and Primary Structure of Cytochrome *f* from the Cyanobacterium, *Plectonema boryanum*
A. Aitken

281 The Amino-Acid Sequence of Trout-Testis Histone H1
A. R. MacLeod, N. C. W. Wong, and G. H. Dixon

293 Oxidation Kinetics of 1,5-Dihydroflavin by Oxygen in Non-aqueous Solvent
V. Favaudon

309 Affinity Chromatography of Glyceraldehyde-3-phosphate Dehydrogenase. A Comparative Study of the Enzymes from Yeast and Sturgeon Muscle
A. F. Chaffotte, C. Roucous, and F. Seydoux

Indexed in Current Contents



SERVA FEINBIOCHEMICA GMBH. & CO.

The EUROPEAN Biochemicals Supply House

D-6900 HEIDELBERG-1 · P.O.B. 105260 · Telefon 06221/12014 · Telex 0461709

Represented in: Austria, Belgium, Danmark, France, Great Britain, Greece, Italy, Netherlands, Norway, Spain, Switzerland, Sweden, Turkey

A 4

European Journal of Biochemistry

This JOURNAL continues the tradition of "Biochemische Zeitschrift" founded in 1906 and edited by C. Neuberg (—Vol. 246), C. Neuberg and W. Grassmann (—Vol. 279), W. Grassmann (—Vol. 317), F. G. Fischer and K. Lang (—Vol. 325), and by a group of leading German biochemists unto the end. Upon completion of Vol. 346 its publication was discontinued in favour of closer European cooperation.

The EUROPEAN JOURNAL OF BIOCHEMISTRY will primarily publish papers on fundamental aspects of biochemistry, molecular biology or cellular biology provided that they describe results which make a sufficient contribution, either experimental or theoretical, to our understanding of biological problems at the chemical or physical level, or that they describe new methods applicable to biochemical problems.

The EUROPEAN JOURNAL OF BIOCHEMISTRY will accept papers from any country—European or not—whether or not the author is a member of the *Federation of European Biochemical Societies (FEBS)*.

The JOURNAL will only accept original papers concisely written and in final form. Negative results will only be accepted when they can be considered to advance our knowledge. Confirmatory information will only be published in exceptional circumstances.

Manuscripts should preferably be written in English, but papers in French or German will also be accepted (with an extended summary in English).

Submission of a paper implies that it presents the results of original research not previously published (except in the form of an abstract or preliminary note or as part of a published lecture, review or thesis), that it is not under consideration for publication elsewhere and that, if accepted, it will not be released by an Agency independent of the author(s), nor be published elsewhere (in the same or in another language) without the written consent of the Editor-in-Chief.

Every author who submits any paper with a view to its publication shall be required to sign an undertaking that submission of the manuscript has been approved by all authors (if this applicable) and by the laboratory in which the work has been performed.

Manuscripts should conform to the "Instructions to Authors" (overleaf) and should be submitted, in triplicate and triple spaced throughout, to:

Professor Claude Liébecq, Editor-in-Chief
European Journal of Biochemistry
Boulevard de la Constitution, 69 (Bte 054)
B-4020 Liège (Belgium)

For each article 50 offprints are supplied free of charge. Additional copies may be ordered at cost price; this must be done at the time when page proofs are returned to the publisher.

Subscription information

Volumes 72—81 (ten volumes each containing two or three issues) will appear in 1977. Information about obtaining back volumes and microform editions available upon request.

Please note that subscribers to the regular printed edition can enter additional subscriptions to this journal in microfiche at a 75 % discount.

All Countries (Except North America). Subscription rate: DM 1320.— plus postage and handling. Orders can either be placed with your bookdealer or sent directly to: Springer-Verlag, Heidelberger Platz 3, D-1000 Berlin 33, W. Germany, Tel. (030) 822001, Telex 01-83319.

North America. Subscription rate: $ 577.30 including postage and handling. Subscriptions are entered with prepayment only. Orders should be addressed to: Springer-Verlag New York Inc., 175 Fifth Avenue, New York, N.Y. 10010, Telex: 0023232235.

The exclusive copyright for all languages and countries, including the right for photomechanical and any other reproductions, also in microform, is vested in the Federation of European Biochemical Societies.

The use of general descriptive names, trade names, trade marks, etc., in this publication, even if the former are not specifically identified, is not to be taken as a sign that such names are exempt from the relevant protective laws and regulations and may accordingly be used freely by anyone.

For subscribers in U.S.A. and Canada second class postage paid at New York. N.Y.
dispatch by Air and Sea Freight Inc. P. O. Box 1305, Long Island City, N. Y. 11101.

Copyright © by the Federation of European Biochemical Societies 1977.

Volume 78 1977

FEBS

European Journal of Biochemistry

Editorial Board

Honorary Chairman: Sir Hans Krebs (Oxford)

Editor-in-Chief: Claude Liébecq (Liège)

Associate Editors: Giorgio Bernardi (Paris), J. Ieuan Harris (Cambridge), Lothar Jaenicke (Köln), Cees Veeger (Wageningen)

Editors: F. Franěk (Praha), G. P. Georgiev (Moskva), H. G. Hers (Bruxelles), T. Keleti (Budapest), M. Lazdunski (Nice) P. Overath (Tübingen), J. G. Reich (Berlin), M. Revel (Rehovot), M. Sela (Rehovot), C. A. Vernon (London), D. H. Williamson (Oxford), H. G. Zachau (München), W. Zillig (Martinsried)

Special Advisor to the Editor-in-Chief: H. Mayer-Kaupp (Heidelberg)

Advisory Board

C. Baglioni (Albany), R. Banerjee (Paris), E. K. F. Bautz (Heidelberg), H. Beaufay (Bruxelles), L. D. Bergelson (Moskva), L. Boross (Szeged), F. Bossa (Camerino), E. M. Bradbury (Portsmouth), R. Braun (Bern), H. Buc (Paris), P. H. W. Butterworth (London), E. Carafoli (Zürich), M. J. Clemens (London), Sir John Cornforth (Brighton), K. Decker (Freiburg i. Br.), G. H. de Haas (Utrecht), W. Doerfler (Köln), J.-P. Ebel (Strasbourg), W. C. Fiers (Gent), C. Gancedo (Madrid), P. B. Garland (Dundee), J.-M. Ghuysen (Liège), D. B. Gower (London), D. O. Hall (London), E. J. M. Helmreich (Würzburg), P. Hemmerich (Konstanz), B. Hirt (Lausanne), E. Hofmann (Leipzig), O. Hoffmann-Ostenhof (Wien), R. Jaenicke (Regensburg), P. Jollès (Paris), M. Ya. Karpeisky (Moskva), E. Katchalski-Katzir (Rehovot), K. Kirschner (Basel), M. Klingenberg (München), H. Klostermeyer (Kiel), G. L. E. Koch, (Cambridge), L. Kovac (Ivánka pri Dunaji), C. G. Kurland (Uppsala), U. G. Lagerkvist (Göteborg), A. Levitzki (Jerusalem), U. Z. Littauer (Rehovot), R. Lontie (Leuven), O. Lüderitz (Freiburg i. Br.), S. Magnusson (Aarhus), C. Milstein (Cambridge), R. Monier (Villejuif), K. Mosbach (Lund), E. A. Newsholme (Oxford), D. Oesterhelt (Würzburg), S. Orrenius (Stockholm), T. A. J. Payens (Ede), D. W. G. Pette (Konstanz), R. J. Planta (Amsterdam), J. Ricard (Marseille), R. Rigler (Stockholm), A. Schellenberger (Halle/Saale), M. C. Scrutton (London), G. Semenza (Zürich), A. Sentenac (Saclay), S. Shaltiel (Rehovot), N. Sharon (Rehovot), D. Shugar (Warszawa), F. Šorm (Praha), F. S. Steven (Manchester), H. Sund (Konstanz), R. Thauer (Marburg/Lahn), R. Thomas (Rhode-St-Genèse), K. F. Tipton (Dublin), A. Trebst (Bochum), D. R. Trentham (Philadelphia), R. Tsanev (Sofiya), V. Ullrich (Homburg), K. van Dam (Amsterdam), D. Vazquez (Madrid), L.-V. von Stedingk (Stockholm), V. P. Whittaker (Göttingen), O. Wieland (München), M. Wiewiorowski (Poznan), L. Wojtczak (Warszawa)

Published by Springer-Verlag Berlin Heidelberg New York
on behalf of the Federation of European Biochemical Societies

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

Eur. J. Biochem. 78, 133–140 (1977)

Structural Rearrangements Due to Ligand Binding and Haem Replacement in Myoglobin and Leghaemoglobins

Nicos A. NICOLA and Sydney J. LEACH
Department of Biochemistry, University of Melbourne

(Received July 30, 1976/June 1, 1977)

Structural rearrangements in sperm whale myoglobin and leghaemoglobins caused by changes in the spin or oxidation state of the iron as a consequence of ligand binding have been measured by difference spectroscopy in the ultraviolet. When compared with the high-spin acetate complex, ligands which cause a transition to the low-spin state also cause large perturbations of tyrosine(s) remote from the haem pocket in myoglobin but only minor perturbations of tryptophan(s) in leghaemoglobin. This may indicate a weaker coupling between events at the haem site and conformational changes in the protein in leghaemoglobins. The absorption spectra of various haem-liganded forms of the two proteins as well as the binding of the dye rose Bengal to the two apoproteins are consistent with weaker interactions between the haem and apoprotein and a more solvent-exposed haem pocket in leghaemoglobin compared with myoglobin.

Leghaemoglobins are plant haemoproteins which resemble animal myoglobins with respect to their molecular weight, monomeric state and their ligand-binding (especially oxygen-binding) properties. Their high oxygen affinity is presumed to be central to their biological function in the root nodules of legumes (see *e.g.* [1]). Most of the reported structural, spectroscopic and ligand-binding studies have been on the major leghaemoglobin component from soybean (*Glycine max*, cr Lincoln). It has recently been shown that this plant produces several leghaemoglobins of differing amino acid sequence but similar overall conformation [1,2], that these characteristics are probably general in *Leguminosae* (Thulborn *et al.*, unpublished) and that the most readily discernable differences between leghaemoglobins of widely different origin and amino acid sequence, is their antigenicity [3]. Comparing the leghaemoglobins with sperm whale myoglobin reveals interesting similarities and differences [2,4]. For the purposes of the present paper, the most pertinent conclusion was that the haem moiety may be more loosely held in the leghaemoglobins than in myoglobin and there is less resistance to changes in spin state on ligand binding in the former case. This is consistent with the observations that soybean leghaemoglobin binds nicotinic acid with much higher affinity [5,6] and its haem moiety is more readily exchangeable [7] than in myoglobin. Evidence has also been adduced that the haem pocket in the plant globin is not as non-polar as that in myoglobin [8,9].

The purpose of the studies described here was to seek evidence of structural changes occurring in both types of haem proteins arising from changes in spin state. The reason for our interest in such changes is the observation by Perutz [10] that high-spin to low-spin transitions in tetrameric haemoglobin are accompanied by deep-seated molecular rearrangements but little is known of corresponding changes in monomeric systems. In order to compare them with myoglobin and to generalise from the results, several leghaemoglobins and a variety of ligands (including nicotinate) and both oxidation states were used. Changes in absorption spectra in the visible, Soret and near-ultraviolet were measured, usually by the difference technique. Finally, the dye rose Bengal was used as a spectroscopic indicator of the polarity of the haem pockets in the two types of proteins.

MATERIALS AND METHODS

Materials

The dye rose Bengal (3′4′5′6′-tetrachloro-2,4,5,7-tetraiodofluorescein disodium salt) was obtained from Fluka. Chlorohemin was lot 8071 from Nutritional Biochemicals Corp. and was recrystallised according to Fischer [11]. Leghaemoglobins were extracted and purified as already described [2,4]. Sperm whale

Abbreviation. CD, circular dichroism.

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

myoglobin was batch 10 from Seravac or batch 52348 from Koch-Light.

Potassium ferricyanide, potassium cyanide, potassium cyanate, sodium acetate, sodium azide, sodium dithionite and sodium fluoride were AR grade from British Drug Houses. Nicotinic acid was AR grade from Townson and Mercer and imidazole from Sigma. Other chemicals were AR grade and glass-distilled water was used throughout.

Concentration Estimations

Absorption spectra for the reduced pyridine haemochromes of soybean, serradella and lupin leghaemoglobins were identical with those for sperm whale myoglobin. Since the spectra for other complexes of these proteins were much more variable than this, the formation of pyridine haemochromes was used as a measure of haemoprotein concentration as in previous studies [2,4].

Formation of Liganded Components

Most liganded complexes of leghaemoglobin and myoglobin were formed simply by titration with the appropriate ligand until no further spectral changes could be observed. The deoxyferrous and oxyferrous complexes were formed as described previously [4] and the percentage of oxyferrous and ferric forms was assayed as described in the same paper.

Formation of Apoproteins

Apoproteins of myoglobin and leghaemoglobins were made from the haem proteins by the acetone/HCl method of Rossi-Fanelli *et al.* [12]. Acetone (BDH) was redistilled from potassium permanganate and Drierite. Glass-distilled water was de-ionised using a Bio-Rad mixed-bed ion-exchange resin. If the removal was carried out at -20 °C the products usually had less than 1% of the original haem content as judged from the A_{400}/A_{280} ratio and were recovered in about 70% yield. Their concentrations and those of the haem proteins were determined as previously described [2,4]. Molar absorption coefficients at 280 nm in phosphate pH 7.5 were 17900 and 16300 M^{-1} cm^{-1} for soybean apoleghaemoglobin and sperm whale apomyoglobin respectively.

Absorption Spectra and Difference Spectra

Absorption spectra were measured with a Cary 14 spectrophotometer. Difference spectra were usually run on this instrument using conditions similar to those reported by Nicola and Leach [13] except that a four-cell arrangement was usually used to allow compensation for the addition of absorbing ligands to the sample cell.

Circular Dichroic Spectra

The instruments and procedures used for measuring CD spectra were those described by Nicola *et al.* [4].

RESULTS

Effect of Spin and Oxidation State on Soret and Visible Absorption Spectra

Spectral data for soybean leghaemoglobin a and the main lupin component have been reported by a number of workers [1,9,14–16] and these have been supplemented by data for the aquoferric, fluoroferric and nicotinoferric as well as the oxyferrous and deoxyferrous soybean derivatives [4]. In the present work, Soret and visible absorption spectra have been measured for the acetate, cyanate, azide and imidazole derivatives of ferric soybean leghaemoglobin. All of these spectra are best discussed by grouping the derivatives as follows: high-spin ferrous = deoxy; low-spin ferrous = oxy, pyridine and nicotinate; high-spin ferric = acetate and fluoride; low-spin ferric = cyanide, imidazole and nicotinate; mixed-spin ferric = aquo, cyanate, azide.

The spectra for soybean leghaemoglobin a are very similar to those reported for lupin leghaemoglobin [9]. They are also quite similar in general form to those for sperm whale myoglobin [17] with corresponding complexes showing the same spin-state tendencies. However, there are differences in the peak absorption intensities and their wavelengths. Table 1 shows that the Soret band for all classes of leghaemoglobin spectra consistently occurs at shorter wavelengths (blue shift) than for corresponding complexes of myoglobin independent of the spin or oxidation state of the iron. Also the mixed-spin complexes for leghaemoglobin appear to be further towards the low-spin state than those for myoglobin (not shown).

Binding of Nicotinic Acid

The binding of nicotinic acid with relatively high affinity to both ferrous and ferric leghaemoglobin is of some interest because this ligand does not bind to myoglobin except at much higher concentrations [18] and because it has been postulated that it may play a regulatory role *in vivo*. This hypothesis and the binding of nicotinic acid to soybean leghaemoglobin a has been discussed in detail by Appleby *et al.* [1,5,6]. We have extended these observations to show that the high-affinity binding of nicotinic acid is probably a general phenomenon for leghaemoglobins. Thus, leg-

Motif Exhibit 1059, 606 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

haemoglobins from serradella, snake bean and lupin all showed changes in absorption spectra for their ferrous forms (reduced with sodium dithionite) when titrated with nicotinic acid in 0.01 M phosphate buffer, pH 7.0, which were qualitatively similar to those of the soybean protein [5,6]. The structural consequences of nicotinic acid binding are best seen using difference spectra. Fig. 1 shows such a set of spectra for the titration of ferric soybean leghaemoglobin a in the near-ultraviolet (Fig. 1A) and visible (Fig. 1B) regions.

The visible difference spectra show the progressive decrease in the charge-transfer band intensities at 494 and 629 nm and the increase in the α and β band intensities at 530 and 556 nm respectively. This is typical for a high-spin to low-spin transition.

The changes which occur concomitantly in the near-ultraviolet, (Fig. 1A), are perhaps more interesting. These changes run parallel with those in the visible and, in fact, some of this difference absorption is due to spectral changes in the haem. Strickland *et al.* [19] however, have shown that neither the haem nor the haem undecapeptide of cytochrome *c* (which contains no aromatic amino acids) display any fine structure bands in the haem absorption region between 250 and 310 nm, even at liquid nitrogen temperatures, so the haem cannot be the origin of the sharp difference peaks seen in Fig. 1(A). The peaks below 280 nm may well arise from the transfer of nicotinic acid from an aqueous environment into the non-polar haem pocket, since the parent spectrum of nicotinic acid shows peaks at 270, 264 and 257 nm. However, the negative difference peak at 294 nm and probably the one at 286 nm must arise from a change in the environment of a tryptophan residue and this change is proportional to the extent of binding of nicotinate. Such negative difference peaks, superimposed on an increasingly positive haem background, suggests that the tryptophan(s) become more exposed to aqueous solvent as the titration proceeds.

Table 1. *Wavelengths of the main Soret peak for various complexes of sperm whale myoglobin, soybean leghaemoglobin a and lupin leghaemoglobin a*

Complex type	Ligand	Wavelength for		
		sperm whale Mb [17]	soybean Lb	lupin Lb [9]
		nm		
Low-spin ferric	CN	423	417	416
High-spin ferric	F	406	403	403
Mixed-spin ferric	H_2O	410	403	404
Low-spin ferrous	O_2	418	412	—
High-spin ferrous	—	434	427	421

'Form' Difference Spectra in the Near-Ultraviolet

'Form' difference spectra (*i.e.* the difference spectra for various liganded complexes against the pure high-spin acetate complex) were recorded in the near-ultraviolet for both sperm whale myoglobin and soybean leghaemoglobin a (see Fig. 2). As mentioned before, haem itself shows no fine structure bands between 250 and 310 nm as evidenced by the difference spectrum for low-spin ferric haem cyanide *versus* high-spin ferric haem chloride. On the other hand, difference

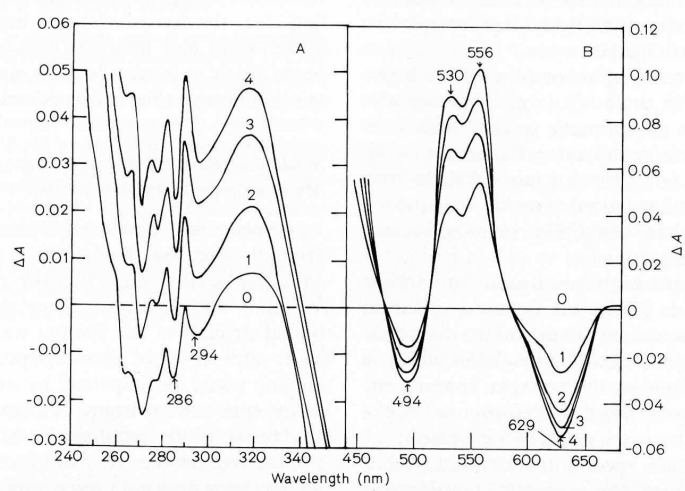


Fig. 1. *Difference spectral titration of soybean ferric leghaemoglobin a with nicotinic acid (A) in the near-ultraviolet and (B) in the visible.* Leghaemoglobin concentration was 33.4 μM in 0.01 M potassium phosphate buffer, pH 5.4, and each addition (1,2,3,4) increased the concentration of nictonic acid by 11.6 μM. Difference spectra were performed with a four-cell arrangement that allowed blanking of both protein and ligand. Saturation was achieved at a final concentration of 58 μM nicotinic acid (not shown)

Motif Exhibit 1059, 607 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

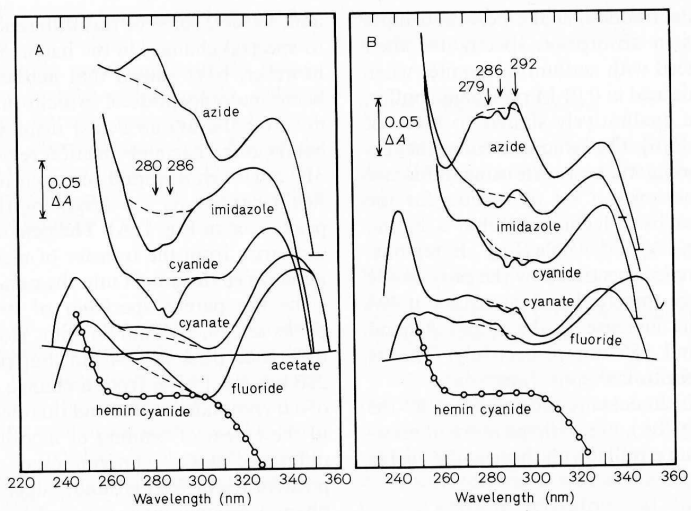


Fig. 2. *Form difference spectra for (A) sperm whale ferric myoglobin and (B) soybean ferric leghaemoglobin a liganded complexes.* For (A) ΔA of 0.05 corresponds to a $\Delta\varepsilon$ of 1.9 mM^{-1} cm^{-1}; for (B) ΔA of 0.05 corresponds to a $\Delta\varepsilon$ of 2.3 mM^{-1} cm^{-1}. All complexes were measured against the corresponding acetate complex as reference. The difference spectra are labelled according to the bound ligand. Some of these difference spectra have been offset vertically for clarity and the baseline position for each marked with a dash through the curve on the long-wavelength side. The dashed curves between 270 and 295 nm are inferred background absorptions due to the haem. For reference, the curve (O——O) for cyanoferric haem (haemin cyanide) *versus* haemin chloride is also shown. All samples were in 0.1 M sodium acetate buffer, pH 5.2, and liganded complexes were formed by titration with solid ligand until no further spectral changes were observed. Leghaemoglobin a concentration was 22 μM and myoglobin concentration 26 μM at 20 °C. A four-cell arrangement was used

spectra of both myoglobin and leghaemoglobin low-spin ferric complexes *versus* the high-spin acetate complex do show fine structure difference bands in this spectral region and these must reflect changes in the environments of the aromatic amino acids upon such a transition. However, there is a difference between the responses of myoglobin and leghaemoglobin to a change in the sixth ligand.

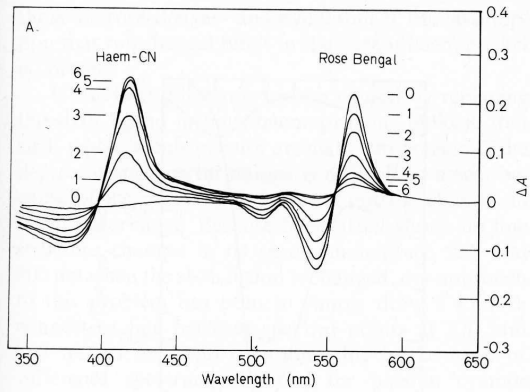
On formation of a low-spin complex from a high-spin one in myoglobin there is a large difference absorption attributable to aromatic amino acids (see especially the cyanide complex *versus* the acetate) with peaks at 280 and 286 nm. These are most likely to arise from the exposure of a buried tyrosine to aqueous solvent since tryptophan would give rise to additional peaks at higher wavelengths.

Replacement of one high-spin ligand for another in myoglobin (fluoride or cyanate for acetate) shows a much smaller aromatic contribution to the difference spectrum. The azide complex of myoglobin which is mixed-spin but predominantly low-spin appears not to follow the expected pattern if the proposed background absorption that we have drawn is correct.

The form difference spectra for leghaemoglobin (Fig. 2B) show much smaller aromatic contributions than do those for myoglobin. The wavelengths of the difference peaks are, however, indicative of a change in the environment of tryptophan rather than of tyrosine, as was observed for the nicotinate *versus* aquoferric complex in Fig. 1A. Since tryptophan difference spectra are expected to be three times larger than for tyrosine, given the same change in environment [13, 20], this means that the change in the tryptophan(s) environment in leghaemoglobins is minor compared to that for the tyrosine(s) in myoglobins. Cyanate (mixed-spin) and fluoride (high-spin) gave aromatic peaks nearly as intense as those given by cyanide (low-spin), imidazole (low-spin) and azide (mixed-spin).

Binding of Rose Bengal to the Apoproteins of Myoglobin and Leghaemoglobin

Coulson and Yonetani [21] have used the anionic, aromatic dye rose Bengal as a sensitizer in photo-oxidation studies on a number of apohemoproteins including apomyoglobin. They showed that this dye bound strongly at one site but weakly at a number of other sites in all of these apoproteins and that this binding could be impaired by addition of stoichiometric amounts of haem. We have used rose Bengal as a probe of the haem pocket in apomyoglobin and apoleghaemoglobin. Its main usefulness for this purpose is that it does not present any of the complications with spin and oxidation states that make it difficult to interpret unambiguously many of the changes that occur in haemoproteins.



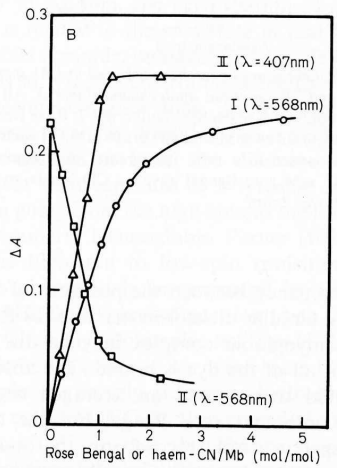


Fig. 3. *Difference spectra produced when soybean apoleghaemoglobin a saturated with rose Bengal is titrated with cyanoferric haem.* (A) The curve labelled 0 is the difference spectrum for leghaemoglobin · rose Bengal *versus* free rose Bengal. Curves 1 – 6 are the difference spectra produced by additions of cyanoferric haem (concentration of haem increasing by 0.9 μM on each addition). Apoleghaemoglobin a was 3.5 μM in sodium bicarbonate (50 mg/l adjusted to pH 7.5). A four-cell arrangement was used. (B) Plot of the binding data in (A). Curve I is a plot of the size of the difference peak at 568 nm when apoleghaemoglobin a is titrated with rose Bengal (representing binding of the dye). Curves II show that when the result of curve I is titrated with cyanoferric haem the difference peak at 568 nm decreases in size while the difference peak at 407 nm (representing haem binding) increases in proportion

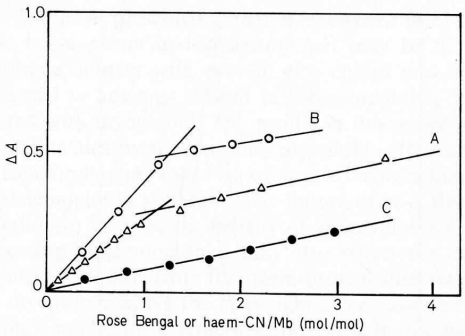


Fig. 4. *Plot of the binding data for rose Bengal and cyanoferric haem to sperm whale apomyoglobin followed by difference spectroscopy.* Curve A is the size of the difference peak at 407 nm when the apoprotein is titrated with cyanoferric haem. Curve B is the size of the difference peak at 568 nm when the apoprotein is titrated with rose Bengal. Curve C is the size of the difference peak at 568 nm when the result of curve A is titrated with rose Bengal. Apomyoglobin concentration was 6.9 μM in 3 ml sodium bicarbonate (50 mg/l, pH 7.5) and consecutive 1, 3 or 6 μl additions of stock rose Bengal (2.91 mM) or cyanoferric haem (2.69 mM) were made each time

In Fig. 3A is shown a typical haem competition experiment after rose Bengal had been bound to soybean apoleghaemoglobin a. The cyano complex of free haem was used because this is known to be monomeric in solution. The difference absorption band due to the binding of cyano-haem is well separated from that due to the bound rose Bengal and it can be clearly seen that as the cyano-haem binds to the protein there is a stoichiometric release of bound rose Bengal. The binding of rose Bengal to apoleghaemoglobin a and the competition with haem is plotted in Fig. 3B. This shows that rose Bengal binds to only one high-affinity site in apoleghaemoglobin a and since it is stoichiometrically released from this site by haem this is probably the haem pocket (though we cannot rule out a strong anticooperative linkage between two different sites).

Fig. 4 shows the results of a similar binding study of rose Bengal to apomyoglobin. In confirmation of the conclusions of Coulson and Yonetani [21] it can be seen that there is one high-affinity site for rose Bengal in apomyoglobin but there are, in addition, a number of weaker binding sites. This is shown by curve (C) which indicates that binding of the dye will still occur even when the pocket is occupied by the haem.

It is clear from these data that, if the rose Bengal concentration is held below that of the apoprotein of either myoglobin or leghaemoglobin, the dye will be bound only at the haem site and the spectral properties of such a complex between 400 and 600 nm will reflect the haem environment. Such spectra are shown for apomyoglobin and apoleghaemoglobin in Fig. 5. The spectrum of rose Bengal is shifted by about 20 nm to higher wavelengths when incorporated into these apoproteins and this is consistent with the transfer from aqueous solution to a non-polar environment as evidenced by the spectral changes that occur when rose Bengal is transferred from water to ethanol. Thermal perturbation difference spectra properties of these complexes are described by Nicola and Leach [8].

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

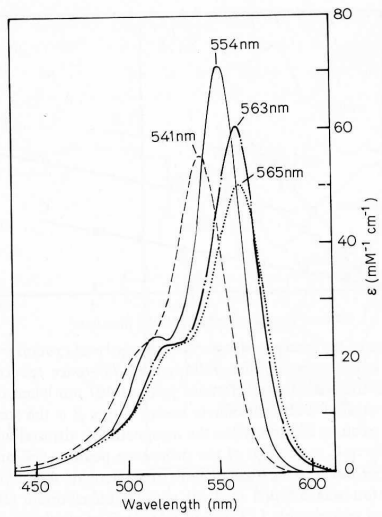


Fig. 5. *Absorption spectra of rose Bengal in (– – – –) water, (——) ethanol, (. . . . .) bound to apomyoglobin and (— · · —) bound to apoleghaemoglobin a.* Peak absorption wavelengths are indicated. For all spectra, rose Bengal concentration was 20 μM. Apomyoglobin and apoleghaemoglobin concentrations were also 20 μM in 0.1 M sodium phosphate buffer, pH 7.0

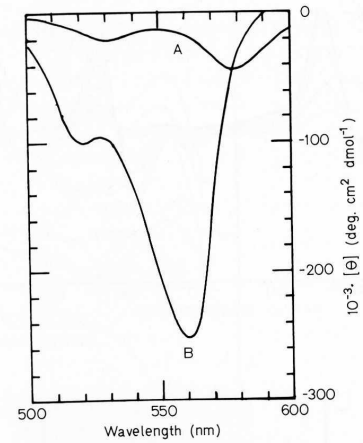


Fig. 6. *Circular dichroism of rose Bengal bound to (A) sperm whale apomyoglobin and (B) soybean apoleghaemoglobin a.* All samples were in 0.1 M potassium phosphate buffer, pH 7. Rose Bengal concentration was (A) 0.104 mM and (B) 0.134 mM. In each case the apoprotein was added in a ratio just greater than equimolar. A solution of 0.15 mM rose Bengal gave no CD bands on its own. Cell pathlength was 2 mm

CD Effects with Rose Bengal

The binding of rose Bengal had no significant effect on the far-ultraviolet CD spectrum of apomyoglobin but increased $[\theta]_{\bar{w}}$ at 220 nm of soybean apoleghaemoglobin from −16300 to −19300 deg. cm^2 dmol^{-1} at pH 7.05, suggesting that the dye stabilises apoleghaemoglobin in much the same way as does the haem group [4].

Fig. 6 shows the CD spectra for rose Bengal bound to apomyoglobin and apoleghaemoglobin a. A solution of rose Bengal free in solution gave no CD peaks but it showed strong negative CD bands when bound to apoleghaemoglobin and much weaker negative CD bands when bound to apomyoglobin. The wavelength positions of the CD peaks in these two complexes show the same trend as seen in corresponding absorption spectra: that is, the peaks are red-shifted on binding. It is not known in which direction the transition moments of rose Bengal point when the dye is bound in the haem pocket but these data are consistent with the postulate that the protein side chains in leghaemoglobin interact with transitions in both the x and y directions of the haem to give negative rotational strengths whereas in myoglobin there is a substantial cancellation of positive and negative strengths so that the CD spectrum can be slightly positive or negative depending on which component predominates [22].

The discrepancy between the positions of the absorption and circular dichroism maxima for the rose-Bengal · apomyoglobin complex may be due to the fact that not all of the dye is bound. The absorption spectra would then record an averaged result for bound and unbound rose Bengal but the circular dichroism spectra would record only the bound dye because the unbound form is optically inactive. Thus, if binding is incomplete, the distinction between the polarity of the leghaemoglobin and the non-polarity of the myoglobin pocket is accentuated because the true absorption maxima would be 578 nm for rose-Bengal · apomyoglobin (from circular dichroism) and 563 nm for rose-Bengal · apoleghaemoglobin (from circular dichroism and absorption). It would, however, decrease the large difference in ellipticity between the rose-Bengal · apomyoglobin and the rose-Bengal · apoleghaemoglobin complexes.

DISCUSSION

The blue shift of leghaemoglobin absorption spectra relative to corresponding myoglobin spectra, by analogy with solvent effects on other chromophores (see *e.g.* [13, 20]), probably indicates a more polar (aqueous) environment of the haem group in leghaemoglobin (see also [8]). This idea is also supported by the blue shift of absorption and CD spectra of the rose-Bengal · apoleghaemoglobin complex relative to

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

those of rose-Bengal · apomyoglobin if our assumption that rose Bengal binds in the vacant haem pocket is correct.

Ultraviolet-difference spectra caused by replacing the sixth ligand in both haemoproteins indicate that such replacements perturb aromatic amino acids. The degree of such perturbations is difficult to assess because of the necessity to subtract out the changes in haem absorbance. Because haem itself shows no fine structure changes in its spectrum between 260 and 300 nm when the sixth ligand is changed, our approach to this problem has been to simply draw a smooth connecting line between spectral points at 270 and 300 nm as an approximation; this is the kind of difference spectrum obtained for haemin cyanide (low-spin) *versus* haemin chloride (high-spin) when protein perturbations are absent (Fig. 2).

For myoglobin, the perturbation of aromatic residues is related to the spin state in that the low-spin complexes (cyanide, imidazole, and azide) show larger changes in tyrosine environment relative to the high-spin acetate complex than do the other high-spin complexes (cyanate, fluoride). For cyanide and imidazole complexes at least, the difference spectra are consistent with exposure of a tyrosine residue (blue shift) in going from the high-spin to the low-spin state. For tetrameric haemoglobin Perutz [10] has shown that the high-spin to low-spin transition occurring upon oxygenation is accompanied by the expulsion of Tyr-H22 from a buried position to an exposed one. Sperm whale myoglobin also has Tyr-H22 in the same position and it seems to serve the same function as in haemoglobin; namely, it forms a hydrogen bond to a peptide carbonyl (FG-5), thus cross-linking the F and H helices [23], so that this change is probably the source of the observed difference spectra.

In contrast to this, the difference spectra for leghaemoglobin show (smaller) perturbations of tryptophan residues upon ligand exchange without the pronounced difference in response between high-spin and low-spin complexes. Even the bulky ligand, nicotinic acid, does not create a significantly larger perturbation to the structure than the other ligands.

There is a high degree of structural homology between myoglobin and leghaemoglobin revealed by sequence alignments [24–29] and the recent X-ray structure of lupin leghaemoglobin [30]. Pursuing this structural homology, neither myoglobin nor leghaemoglobin contain tyrosine or tryptophan residues in or very near to the haem pocket and this implies that ligand exchange and especially a high-spin to low-spin transition causes rather extensive structural rearrangements through the protein which involve these residues. These rearrangements may be more significant in myoglobin than in leghaemoglobin since the perturbation of the tyrosine residue is much greater than that of tryptophan in leghaemoglobin.

We have presented evidence elsewhere [4,8] that the haem group in leghaemoglobin may be in less intimate contact with protein side chains and more exposed to aqueous solvent in leghaemoglobin compared with myoglobin. The results in this paper support this idea by showing that myoglobin structure is affected relatively more by spin-state changes than is leghaemoglobin structure. The failure of rose Bengal binding to increase the helicity of apomyglobin (as it does for leghaemoglobin) may also reflect the more stringent requirements for haem-protein interactions in myoglobin where the fit is tighter (*i.e.* rose Bengal cannot mimic the interactions provided by the haem in native myoglobin but can in the more flexible leghaemoglobin pocket).

The concept of a more flexible, open haem pocket in leghaemoglobin helps to explain many of the properties of leghaemoglobin compared with myoglobin, in particular its ability to bind bulky ligands and its high oxygen affinity. The absence of Tyr-H22 in leghaemoglobin also means that a tyrosine is not transferred from a non-polar to an aqueous environment (an energy-requiring process) upon oxygenation (high-spin to low-spin transition) further accounting for the higher oxygen affinity of leghaemoglobin compared with myoglobin.

The authors are grateful to Dr C. A. Appleby (Division of Plant Industry, Commonwealth Scientific and Industrial Research Organization, Canberra) for gifts of soybean leghaemoglobin a and for helpful discussions. They also wish to thank Professor M. J. Dilworth (School of Environmental and Life Sciences, Murdoch University, Murdoch, W.A.) and Dr W. J. Broughton (School of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia) for valuable gifts of serradella and lupin leghaemoglobins. This work was carried out with the aid of a grant from the Australian Research Grants Committee to S.J.L. N.A.N. was supported by a Commonwealth Scientific and Industrial Research Organisation Studentship.

REFERENCES

1. Appleby, C. A. (1974) in *The Biology of Nitrogen Fixation* (Quispel, A., ed.) ch. 11.5, pp. 522–554, North Holland Publishing Co., Amsterdam.
2. Appleby, C. A., Nicola, N. A., Hurrell, J. G. R. & Leach, S. J. (1975) *Biochemistry*, *14*, 4444–4450.
3. Hurrell, J. G. R., Nicola, N. A., Broughton, W. J., Dilworth, M. J., Minasian, E. & Leach, S. J. (1976) *Eur. J. Biochem. 66*, 389–399.
4. Nicola, N. A., Minasian, E., Appleby, C. A. & Leach, S. J. (1975) *Biochemistry, 14*, 4151–5149.
5. Appleby, C. A., Wittenberg, B. A. & Wittenberg, J. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 564–568.
6. Appleby, C. A., Wittenberg, B. A. & Wittenberg, J. B. (1973) *J. Biol. Chem. 248*, 3183–3187.
7. Ellfolk, N., Pertila, U. & Sievers, G. (1973) *Acta Chem. Scand. 27*, 3601–3602.
8. Nicola, N. A. & Leach, S. J. (1977) *Biochemistry, 16*, 50–58.
9. Atanasov, B. P. & Zhiznevskaya, G. Ya (1975) *Mol. Biol. 9*, 491–501.

Motif Exhibit 1059, 611 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

10. Perutz, M. F. (1970) *Nature (Lond.) 228*, 726–734.
11. Fischer, H. (1941) *Organic Syntheses, 21*, 53–55.
12. Rossi-Fanelli, A., Antonini, E. & Caputo, A. (1958) *Biochim. Biophys. Acta, 30*, 608–615.
13. Nicola, N. A. & Leach, S. J. (1976) *Int. J. Peptide Protein Res. 8*, 393–415.
14. Ellfolk, N. (1961) *Acta Chem. Scand. 15*, 975–984.
15. Sternberg, H. & Virtanen, A. I. (1952) *Acta Chem. Scand. 6*, 1342–1352.
16. Appleby, C. A. (1969) *Biochim. Biophys. Acta, 188*, 222–229.
17. Antonini, E. & Brunori, M. (1971) *Hemoglobin and Myoglobin in Their Reactions with Ligands*, p. 436, North Holland Publishing Co., Amsterdam.
18. Keilin, J. (1966) in *Hemes and Hemoproteins*, Chance, B., Estabrook R. W. & Yonetani, T., eds) pp. 171–191, Academic Press, New York.
19. Strickland, E. H., Horwitz, J., Kay, E., Shannon, L. M., Wilchek, M. & Billups, C. (1971) *Biochemistry, 10*, 2631–2638.
20. Herskovits, T. T. (1967) *Methods Enzymol. 11*, 748–775.
21. Coulson, A. F. W. & Yonetani, T. (1972) *Eur. J. Biochem. 26*, 125–131.
22. Hsu, M.-C. & Woody, R. W. (1971) *J. Am. Chem. Soc. 93*, 3515–3525.
23. Kendrew, J. C. (1962) *Brookhaven Symp. Biol. 15*, 216–228.
24. Ellfolk, N. & Sievers, G. (1971) *Acta Chem. Scand. 25*, 3532–3534.
25. Dayhoff, M. O. (ed.) (1972) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington.
26. Ellfolk, N. & Sievers, G. (1974) *Acta Chem. Scand. B28*, 1245–1246.
27. Richardson, M., Dilworth, M. J. & Scawen, M. D. (1975) *FEBS Lett. 51*, 33–37.
28. Lehtovaara, P. & Ellfolk, N. (1975) *Eur. J. Biochem. 54*, 577–584.
29. Nicola, N. A. (1975) Ph. D. Thesis, Melbourne University.
30. Vainshtein, B. K., Harutyunyan, E. H., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavolovsky, A. G., Grebenko, A. I. & Konareva, N. V. (1975) *Nature (Lond.) 254*, 163–164.

N. A. Nicola, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia 3052

S. J. Leach*, The Russell Grimwade School of Biochemistry, University of Melbourne, Parkville, Victoria, Australia 3052

* To whom correspondence should be addressed

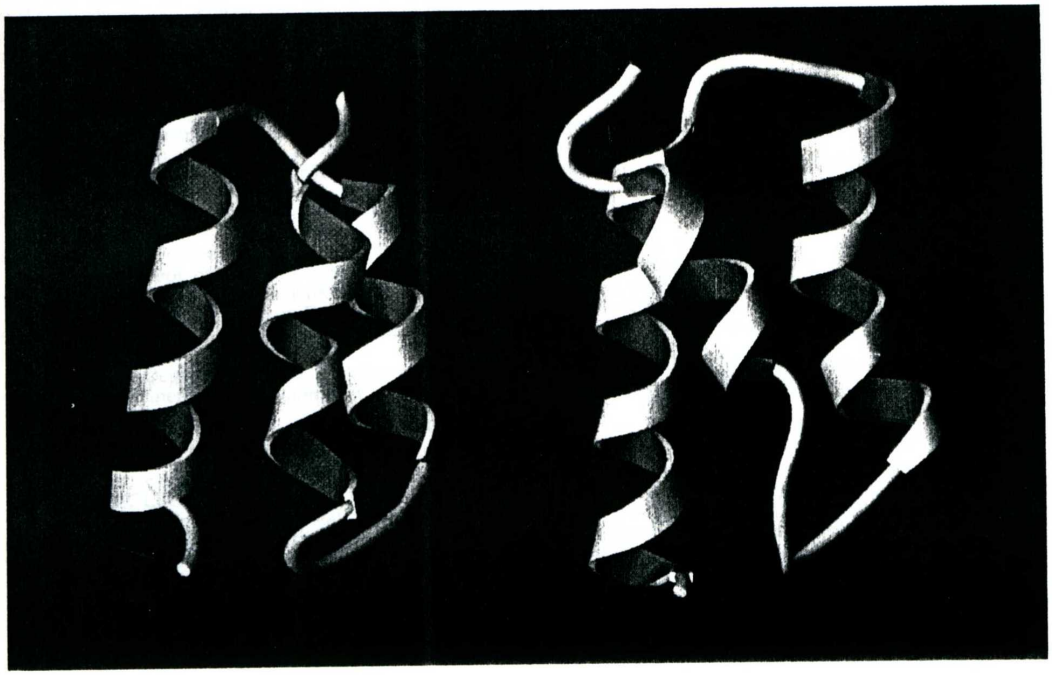
Motif Exhibit 1059, 612 of 653

Ex. 1045 Hargrove 1997

Volume 266
Number 5
14 March 1997

JMB

JOURNAL OF MOLECULAR BIOLOGY



DEAL

ACADEMIC PRESS

JMB

LINDA HALL
MAR 26 1997
LIBRARY

Editor-in-Chief
Peter Wright
Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, U.S.A.

Executive Editor
Jonathan Karn
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

Founding Editor
Sir John Kendrew

Consulting Editor
Sydney Brenner

Advisory Editors
Fred Cohen
Alan R. Fersht
Max Gottesman
Robert Huber
Aaron Klug
Moshe Yaniv

Board of Editors
Wolfgang Baumeister, Max-Planck-Institut für Biochemie, Am klopferspitz 18a, D-82152 Martinsried bei Munchen, Germany
Nam-Hai Chua, Laboratory of Plant Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, U.S.A.
Fred Cohen, Dept of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94143-0450, U.S.A.
David E. Draper, Dept of Chemistry, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, U.S.A.
Richard H. Ebright, The State University of New Jersey, Rutgers, Waksman Institute, PO Box 759, Piscataway, NJ 08855-0759, U.S.A.
Alan R. Fersht, University Chemical Laboratory, Cambridge University, Lensfield Road, Cambridge, CB2 1EW, U.K.
Max Gottesman, Institute of Cancer Research, College of Physicians and Surgeons of Columbia University, 701 W. 168th Street, New York, NY 10032, U.S.A.
Barry Holland, Institut de Genetique et Microbiologie, Batiment 409, Universite de Paris XI, 91405 Orsay Cedex 05, France
Barry Honig, Dept of Biochemistry & Molecular Biophysics, College of Physicians & Surgeons of Columbia University, 630 W. 168th Street, NY 10032, U.S.A.
Robert Huber, Max Planck Institute für Biochimie, 82152 Martinsried bei Munchen, Germany
Aaron Klug, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.
Jeffrey H. Miller, Dept of Microbiology, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024, U.S.A.
Michael F. Moody, School of Pharmacy, University of London, 29–39 Brunswick Square, London WC1N 1AX, U.K.
Kiyoshi Nagai, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.
Steven Reed, Dept of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, U.S.A.
Douglas Rees, Division of Chemistry and Chemical Engineering, California Institute of Technology, 147-75CH, Pasadena, CA 91125, U.S.A.
Timothy J. Richmond, Institut für Molekularbiologie und Biophysik Eidgenössische Technische Hochschule, Hönggerberg, CH 8093 Zürich, Switzerland
Gerald Smith, Fred Hutchinson Cancer Research Center, Basic Sciences Division - A1-162, 1124 Columbia Street, Seattle, WA 98104-2092, U.S.A.
Janet M. Thornton, Biomolecular Structure and Modelling Unit, Dept of Biochemistry and Molecular Biology, University College of London, Room 628, Gower Street, London WC1E 6BT, U.K.
Ignacio Tinoco, Dept of Chemistry, University of California, Berkeley, CA 94720-1460, U.S.A.
Gunnar Von Heijne, Dept. of Biochemistry, University of Stockholm, S-106 91 Stockholm, Sweden
James Wells, Protein Engineering, Genentech Inc, 460 Point San Bruno Blvd, South San Francisco, CA 94080, U.S.A.
Ian Wilson, Dept of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, U.S.A.
Keith R. Yamamoto, Dept of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94143-0450, U.S.A.
Mitsuhiro Yanagida, Dept of Biophysics, Faculty of Science, Kyoto University, Sakyo-Ku, Kyoto 606, Japan
Moshe Yaniv, Department of Biotechnology, Pasteur Institute, 25 rue de Dr Roux, 75724 Paris Cedex 15, France

Editorial Office
Journal of Molecular Biology, Academic Press Editorial Office, 10d St Edward's Passage, Cambridge CB2 3PJ, U.K.

JOURNAL OF MOLECULAR BIOLOGY: ISSN 0022–2836. Volumes 265–274, 1997, is published weekly (except the last two weeks in December; 50 issues in total) by Academic Press at 24–28 Oval Road, London NW1 7DX, England. Annual worldwide subscription price including postage: £2280. Personal subscription rates are available from the publisher. Subscription orders should be sent to Academic Press, Foots Cray High Street, Sidcup, Kent DA14 5HP, U.K. (Tel: 0181–300 3322). Send notices of changes of address to the publisher at least 6–8 weeks in advance, including both old and new addresses.

Academic Press journals do not levy page charges.

Buyers in Canada should add GST at the current rate of 7%. Periodicals postage paid at Jamaica, NY 11431, U.S.A.

Air freight and mailing in the U.S.A. by Publications Expediting Inc., 200 Meacham Avenue, Elmont, NY 11003, U.S.A.

U.S.A. POSTMASTERS: send change of addresses to JOURNAL OF MOLECULAR BIOLOGY, c/o Publications Expediting, Inc., 200 Meacham Avenue, Elmont, NY 11003, U.S.A.

Printed in U.K., on acid free paper.

Listed in Biological Abstracts, Science Citation Index, Chemical Abstracts, Reference Update, Excerpta Medica database EMBASE, Current Contents, SCISEARCH, Research Alert, Biotechnology Citation Index, Biochemistry and Biophysics Citation Index, Index Medicus, Genetics Abstracts, Biological & Agricultural Index.

JMB

Volume 266, Number 5

Contents

Communication

Solution Structure of the Albumin-binding GA Module: A Versatile Bacterial Protein Domain (mb960856) — M. U. Johansson, M. de Château, M. Wikström, S. Forsen, T. Drakenberg and L. Björck — 859–865

Articles

Design of CytR Regulated, cAMP-CRP Dependent Class II Promoters in *Escherichia coli*: RNA Polymerase-Promoter Interactions Modulate the Efficiency of CytR Repression (mb960852) — H.-H. Kristensen, P. Valentin-Hansen and L. Søgaard-Andersen — 866–876

Analysis of Interactions Between the Codon–Anticodon Duplexes within the Ribosome: Their Role in Translation (mb960802) — V. I. Lim — 877–890

Protein P7 of Phage ϕ6 RNA Polymerase Complex, Acquiring of RNA Packaging Activity by *In Vitro* Assembly of the Purified Protein onto Deficient Particles (mb960817) — J. T. Juuti and D. H. Bamford — 891–900

The Bacteriophage ø29 Packaging Proteins Supercoil the DNA Ends (mb960843) — S. Grimes and D. Anderson — 901–914

RNA-DNA Hybrid Formation at a Bacteriophage T4 Replication Origin (mb960844) — K. Carles-Kinch and K. N. Kreuzer — 915–926

Single-stranded DNA Binding Properties of the uvsY Recombination Protein of Bacteriophage T4 (mb960829) — M. A. Sweezy and S. W. Morrical — 927–938

A Survey of Polypeptide Deformylase Function Throughout the Eubacterial Lineage (mb960835) — D. Mazel, E. Coïc, S. Blanchard, W. Saurin and P. Marlière — 939–949

A Dynamic *in Vivo* View of the HIV-I Rev-RRE Interaction (mb960858) — B. Charpentier, F. Stutz and M. Rosbash — 950–962

Crystal Structure of the Gramicidin/Potassium Thiocyanate Complex (mb960837) — D. A. Doyle and B. A Wallace — 963–977

Module Swaps between Related Translocator Proteins pIV^{f1}, pIV^{IKe} and PuID: Identification of a Specificity Domain (mb960866) — S. Daefler, M. Russel and P. Model — 978–992

Traffic Control of Completely Assembled MHC Class I Molecules Beyond the Endoplasmic Reticulum (mb960822)	**S. Joyce**	993–1001
Three-dimensional Reconstruction of *Limulus polyphemus* Hemocyanin from Cryoelectron Microscopy (mb960840)	**J.-C. Taveau, N. Boisset, J. Lamy, O. Lambert and J. N. Lamy**	1002–1015
Crystal Structures of *Escherichia coli* and *Salmonella typhimurium* 3-Isopropylmalate Dehydrogenase and Comparison with their Thermophilic Counterpart from *Thermus thermophilus* (mb960797)	**G. Wallon, G. Kryger, S. T. Lovett T. Oshima, D. Ringe, G. A. Petsko**	1016–1031
Characterization of Recombinant Soybean Leghemoglobin *a* and Apolar Distal Histidine Mutants (mb960833)	**M. S. Hargrove, J. K. Barry E. A. Brucker, M. B. Berry G. N. Phillips Jr, J. S. Olson R. Arredondo-Peter, J. M. Dean R. V. Klucas and G. Sarath**	1032–1042
Structure-based Identification of a Ricin Inhibitor (mb960865)	**X. Yan, T. Hollis, M. Svinth, P. Day A. F. Monzingo, G. W. A. Milne and J. D. Robertus**	1043–1049
Energetics of Target Peptide Recognition by Calmodulin: A Calorimetric Study (mb960785)	**P. L. Wintrode and P. L. Privalov**	1050–1062
Author Index		1063–1064

The home page for the Journal of Molecular Biology can be found at

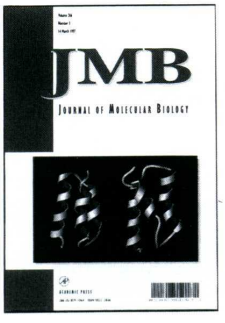
http://www.hbuk.co.uk/ap/journals/mb.htm

IDEAL, Academic Press' online scientific journal library, where guest users can freely search and browse journal abstracts, can be found at:

http://www.idealibrary.com/ (San Jose, USA) and
http://www.europe.idealibrary.com/ (Bath, UK)

The main Academic Press home page is at http://www.apnet.com/ where you will find general information about Academic Press and the AP Online Book Catalog.

Copyright © 1997 Academic Press Limited. The appearance of the code at the bottom of the first page of a paper in this Journal indicates the copyright owner's consent that copies of the paper may be made for personal or internal use, or for the personal or internal use of specific clients in the U.S.A. This consent is given on the condition, within the U.S.A., that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970, for copying beyond that permitted by Section 107 or 108 of the U.S. Copyright Law. This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, for resale or for copying or distributing copies outside the U.S.A.



Cover Illustration: Solution structure of the albumin-binding GA molecule—a versatile bacterial protein domain. See paper by Johansson *et al. J. Mol. Biol.* **266,** 859–865.

Typeset by BPC Digital Techset Ltd; printed and bound by BPC Wheatons Ltd
Both members of The British Printing Company Ltd
This journal is printed on acid-free paper

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

J. Mol. Biol. (1997) **266**, 1032–1042

JMB

Characterization of Recombinant Soybean Leghemoglobin *a* and Apolar Distal Histidine Mutants

Mark S. Hargrove[1]*, Jennifer K. Barry[1], Eric Allen Brucker[1] Michael B. Berry[1], George N. Phillips Jr[1], John S. Olson[1] Raúl Arredondo-Peter[2], Jeanenne M. Dean[2], Robert V. Klucas[2] and Gautam Sarath[2]

[1]*Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77005-1892, USA*

[2]*Department of Biochemistry University of Nebraska-Lincoln Lincoln, NE 68588-0664, USA*

**Corresponding author*

The cDNA for soybean leghemoglobin *a* (Lb*a*) was cloned from a root nodule cDNA library and expressed in *Escherichia coli*. The crystal structure of the ferric acetate complex of recombinant wild-type Lb*a* was determined at a resolution of 2.2 Å. Rate constants for O_2, CO and NO binding to recombinant Lb*a* are identical with those of native soybean Lb*a*. Rate constants for hemin dissociation and auto-oxidation of wild-type Lb*a* were compared with those of sperm whale myoglobin. At 37°C and pH 7, soybean Lb*a* is much less stable than sperm whale myoglobin due both to a fourfold higher rate of auto-oxidation and to a ~600-fold lower affinity for hemin. The role of His61(E7) in regulating oxygen binding was examined by site-directed mutagenesis. Replacement of His(E7) with Ala, Val or Leu causes little change in the equilibrium constant for O_2 binding to soybean Lb*a*, whereas the same mutations in sperm whale myoglobin cause 50 to 100-fold decreases in K_{O_2}. These results show that, at neutral pH, hydrogen bonding with His(E7) is much less important in regulating O_2 binding to the soybean protein. The His(E7) to Phe mutation does cause a significant decrease in K_{O_2} for Lb*a*, apparently due to steric hindrance of the bound ligand. The rate constants for O_2 dissociation from wild-type and native Lb*a* decrease significantly with decreasing pH. In contrast, the O_2 dissociation rate constants for mutants with apolar E7 residues are independent of pH, suggesting that hydrogen bonding to the distal histidine residue in the native protein is enhanced under acid conditions. All of these results support the hypothesis that the high affinity of Lb*a* for oxygen and other ligands is determined primarily by enhanced accessibility and reactivity of the heme group.

© 1997 Academic Press Limited

Keywords: soybean leghemoglobin; ligand binding; site-directed mutagenesis

Introduction

Leghemoglobins (Lb) are small monomeric heme proteins found in root nodules of leguminous plants. These proteins have unusually high affinities for oxygen compared to vertebrate myoglobins and hemoglobins. The association rate constants for O_2 binding are very large (1×10^8 to $3 \times 10^8\ M^{-1}\ s^{-1}$), whereas the dissociation rate constants are in the "normal" range for hemoglobins (5 to $30\ s^{-1}$; Gibson *et al.*, 1989). High oxygen affinity is required to prevent inhibition of nitrogen fixation, and a moderate dissociation rate constant is necessary for facilitating diffusion of oxygen to a high-affinity bacteroid oxidase (Appleby, 1992).

The tertiary folding patterns of lupin and soybean Lb are very similar to those of mammalian myoglobins, and the α and β subunits of human hemoglobin. This pattern consists of seven α-helices (A to H, minus the D helix found in most mammalian myoglobins) wrapped around an iron-protoporphyrin IX prosthetic group that is coordinated to the histidine side-chain of the eighth residue in the F helix. Leghemoglobins also contain a "distal" histidine residue at position E7. Polar interactions involving His-(E7) in mammalian myoglobins have

Abbreviations used: Lb, leghemoglobin; EPR, electron paramagnetic resonance; Mb, myoglobin.

0022–2836/97/101032–11 $25.00/0/mb960833

© 1997 Academic Press Limited

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

been shown to play a key role in regulating ligand binding (Springer *et al.*, 1994). The evidence for hydrogen bonding between bound O_2 and the distal histidine residue in leghemoglobin is less clear.

Comparisons of the ligand binding properties of leghemoglobins with those from a wide variety of animal myoglobins and hemoglobins have suggested that the small rates of O_2 dissociation from soybean and lupin Lb are due to hydrogen bonding with His(E7) (Appleby *et al.*, 1983; Mims *et al.*, 1983; Stetzkowski *et al.*, 1979; Rohlfs *et al.*, 1988). Initial electron paramagnetic resonance (EPR) experiments with cobalt-containing soybean Lb indicated weak hydrogen bonding at neutral pH and a significant strengthening of this interaction at acid pH (Ikeda-Saito *et al.*, 1981). More recent spin echo EPR studies confirmed this observation and further suggested that hydrogen bonding with the distal histidine residue may occur only at low pH (Lee *et al.*, 1993). Fuchsman & Appleby (1979) observed a single narrow IR band at $\nu_{C\text{-}O} = 1947\ cm^{-1}$ for Lb*a*CO at pH 7. This peak corresponds to the main band observed in sperm whale MbCO ($\nu_{C\text{-}O} = 1945\ cm^{-1}$) and indicates a similar electrostatic environment around the bound ligand in both proteins. More recently, Harutyunyan *et al.* (1995) reported a distance of 2.7 Å between the second ligand atom and N^{ε} of His(E7) in lupin oxyleghemoglobin II, indicating a hydrogen bond between the imidazole group and bound O_2. Thus, hydrogen bonding is thought to occur in leghemoglobins, especially at low pH. However, the strength of this interaction and its exact role in governing the kinetics and equilibria of ligand binding to plant hemoglobins have not been established directly by site-directed mutagenesis.

Like all oxygen binding proteins, leghemoglobins must be in the Fe(II) oxidation state to bind oxygen. Within the root nodule, the protein is kept in the ferrous state by a reductase that catalyses the reduction of metleghemoglobin by NADH (Saari & Klucas, 1984; Ji *et al.*, 1994a,b). However, little is known about the molecular recognition and protein-protein interaction between the enzyme and its substrate. Soybean root nodules contain four leghemoglobin gene products, termed Lba, Lbc1, Lbc2 and Lbc3, and four variants of these four proteins in which the N termini are acetylated (Appleby, 1984). The amino acid sequences and ligand binding properties of the four gene products are very similar (Martin *et al.*, 1990). Several studies have shown changes in the relative concentrations of the individual leghemoglobins during plant growth, but the physiological consequences of differential expression and turnover are not well understood (Appleby, 1984).

The expression of recombinant hemoglobins and myoglobins has allowed systematic studies of the structural factors that regulate ligand binding to these proteins (Nagai *et al.*, 1987; Mathews *et al.*, 1991; Kloeck *et al.*, 1993; Springer *et al.*, 1994). In order to carry out similar mechanistic studies, we have cloned Lb*a* from a soybean cDNA library, inserted it into two expression vectors, and used both constitutive and inducible systems to express large quantities of protein. The crystal structure of recombinant wild-type Lb*a* was determined to 2.2 Å and compared with that of the native protein. The ligand binding, spectral, auto-oxidation and hemin dissociation properties of native and recombinant Lb*a* were also compared in detail. His61(E7) to Ala, Val, Leu and Phe mutations were constructed to examine the role of the distal histidine residue in governing O_2 binding. Finally, Tyr30(B10) to Phe and Leu65(E11) to Val mutations were made to examine the structural origin of differences between the rates of O_2 binding to lupin LbII and soybean Lb*a*.

Results

Properties of wild-type Lb*a*

As described in Materials and Methods, the N-terminal methionine residue of recombinant Lb*a* is removed post-translationally during constitutive expression in *Escherichia coli* TB1 cells, making the wild type recombinant Lb*a* chemically identical with native Lb*a* (Figure 1). The absorbance spectra of wild-type metLb, deoxyLb, LbO_2 and LbCO are identical with those of the native protein (Table 1). The kinetic parameters for O_2, CO and NO binding, hemin loss and auto-oxidation are also the same for the wild-type and native proteins (Tables 2, 4, and 5). These data show that the protein obtained from *E. coli* TB1 cells is functionally and spectrally the same as native soybean leghemoglobin *a*. Lbs expressed in the inducible pET28a/BL21DE3 system are only partially processed, yielding a mixed sample of N-met and "native-like" protein. However, the samples prepared from the pET28a/BL21DE3 system are kinetically and spectrally identical with those prepared in TB1. Furthermore, it has been observed that addition of an entire N-terminal S-tag (Novagen) domain has no effect on the kinetic or spectral properties of the Lbs examined here (unpublished results).

Crystal structure of wild type acetomet-Lb*a*

The crystallization system described by Vainshtein *et al.* (1975) was followed starting with aquomet wild-type Lb*a*. Since 0.5 M ammonium acetate is present in the solution, the acetomet form of the protein is produced in the crystals. The backbone and side-chain positions in the acetate complex of wild-type metLb*a* are very similar to those of the nicotinate complex of native metLb*a* (Ollis *et al.*, 1983; Ellis *et al.*, 1997). An overlay of the backbone atoms plus the side-chains of the proximal and distal histidine residues of the two structures is shown in Figure 2. The native nicotinate structure is shown in yellow and molecule a of the two molecules in the asymmetric unit of the wild-type structure is shown in blue. The ligands have been

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

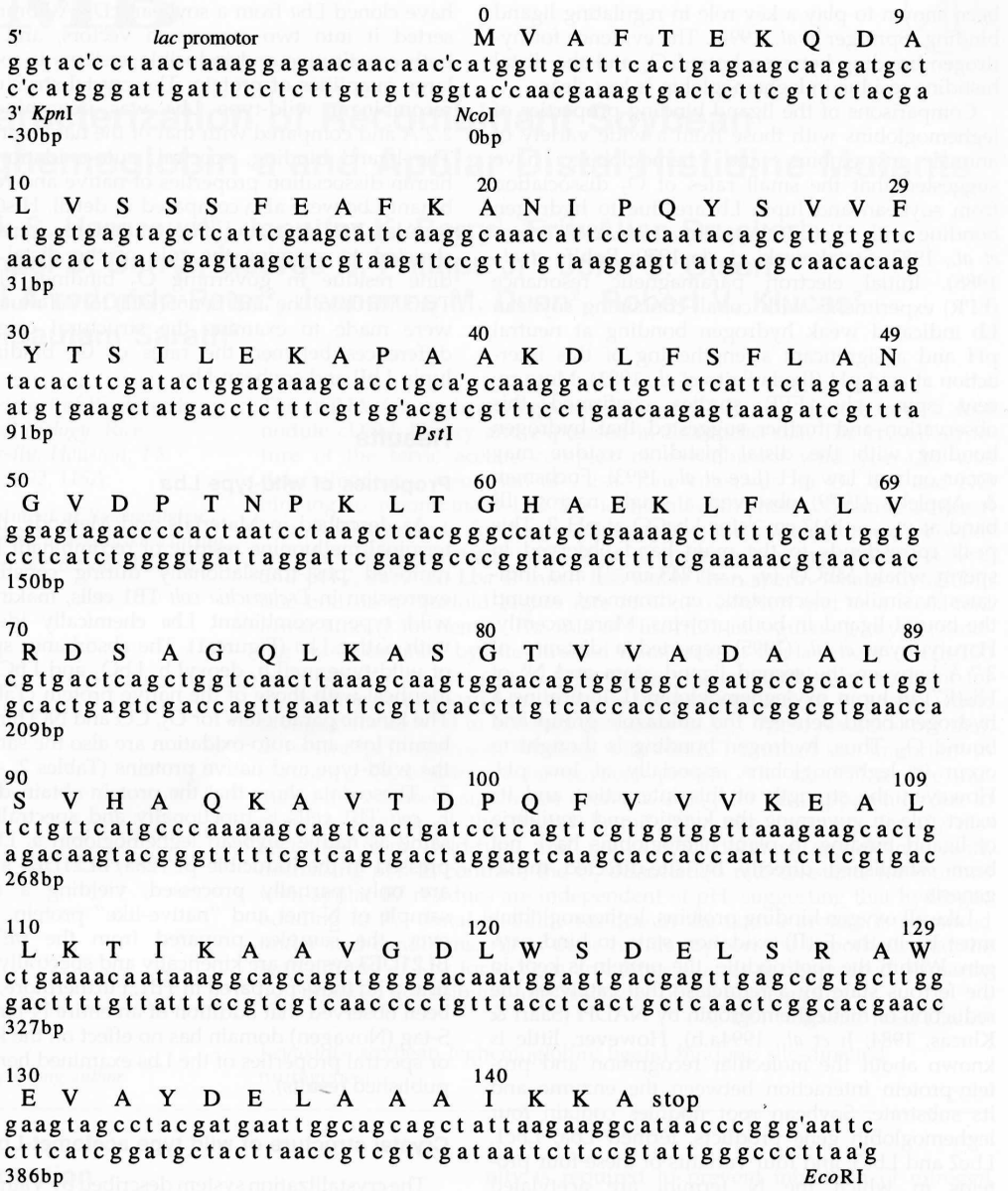


Figure 1. The cDNA and amino acid sequence of recombinant soybean Lb*a* in the pEMBL19$^+$ plasmid pLb*a*. The ribosome binding site lies between the 5′ *Kpn*I and *Nco*I restriction sites. The flanking 5′ *Nco*I and 3′ *Eco*RI restriction sites can be used to move the gene independently of the promoter region. The gene sequence corresponds with that observed for soybean Lb*a* by Hyldig-Nielsen *et al.* (1982).

removed for clarity. These data show that the protein obtained from *E. coli* is structurally the same as native soybean Lb*a*. The most significant differences between the two structures are in the positions of the heme group, distal histidine residue, portions of the CD and FG corners, and part of the F helix including His(F8). These differences appear to be due to the size of the coordinating ligand. In the native structure, the large nicotinate ring forces the distal histidine residue out toward solvent and pushes the heme group downward. The proximal histidine residue, F helix and FG corner rotate in order to compensate for this movement of the porphyrin ring. The CD corner also appears to move away from the bound ligand, but this region is either mobile or disordered in both structures.

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Table 1. Comparison of the wavelength maxima for the Soret absorbance bands of soybean Lb and sperm whale Mb at pH 7, 25°C

Protein	Wavelength peak (nm)				
	Fe(II)	Fe(III)	$Fe(II)O_2$	Fe(II)CO	Fe(II)NO
Native Lb	428	403	409	416	418
Wild-type Lb	428	403	409	416	418
H61L Lb	425	395	414	419	418
Wild-type Mb	434	409	415	424	421
H64L Mb	434	393	415	425	421

Figure 3 shows electron density around the heme binding site of wild-type Lb*a*. The view is looking into the pocket from solvent with the distal portion of the heme pocket located in the upper half of the drawing. The heme is shown in white, the acetate ligand in yellow and the protein chain in red. Well-resolved density for His61(E7) is seen next to the bound acetate group. Other prominent residues are (counter-clockwise): Leu65(E11), located just behind His61(E7); and Phe44(CD1), located to the left of His61(E7), just above the heme plane.

The distal histidine residue in the structure of native soybean metLb*a* appears to be hydrogen bonding with the carbonyl oxygen atom of the nicotinate ligand (Ollis *et al.*, 1983). In the wild-type metLb*a* structure, N^{ε} of His61(E7) is 2.64 Å and 3.08 Å away from the carbonyl oxygen atom of the acetate group in the two protein molecules found in the asymmetric unit. This proximity suggests that a hydrogen bond is also formed between bound acetate and the distal histidine residue in the wild-type protein. Harutyunyan *et al.* (1995) have shown that a hydrogen bond occurs between the distal histidine residue and the second coordinated ligand atom in the crystal structure of lupin LbO_2. Thus, in leghemoglobins, the distal histidine residue appears to be quite flexible and is capable of interacting electrostatically with ligands of varying size.

Role of His61(E7) in O_2 binding

Hydrogen bonding between bound oxygen and the distal histidine residue in sperm whale myoglobin has been established by X-ray crystallography, and its effect on oxygen affinity has been determined in site-directed mutagenesis studies (Springer *et al.*, 1994). Following the strategy of Springer *et al.* (1989), we replaced His61(E7) with Ala, Val, Leu and Phe to examine quantitatively how oxygen binding is affected by loss of hydrogen-bonding potential at this position in Lb*a*. Table 2 presents a comparison of the effects of these E7 mutations on the rate and equilibrium constants for O_2 binding to soybean Lb*a* with those observed previously for O_2 binding to sperm whale myoglobin.

In general, the aliphatic replacements have a much smaller effect on O_2 binding to leghemoglobin than to myoglobin. These differences are particularly striking when the oxygen dissociation rate constants are examined. Only the Phe-E7 mutation in

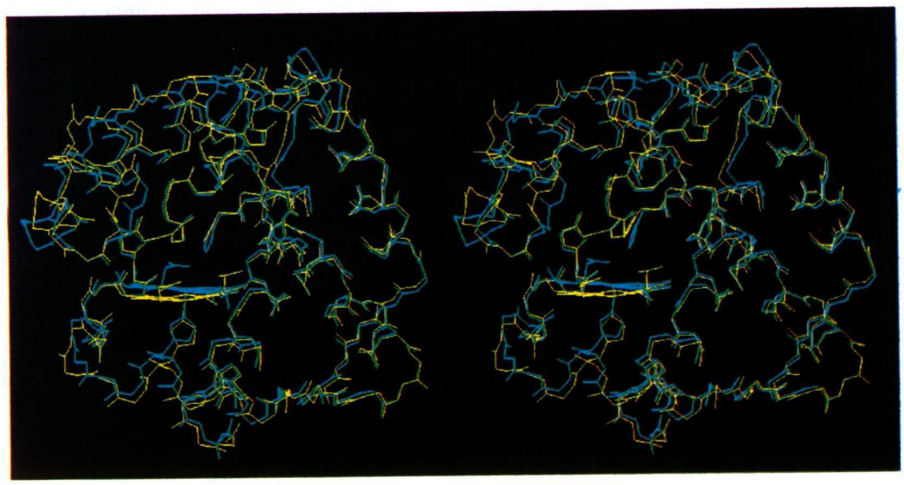


Figure 2. Crystal structure of the acetate complex of recombinant soybean metLb*a*. A stereo overlay of the backbone atoms of the nicotinate complex of native metLb*a* (yellow) and wild-type metLb*a* (blue) including the proximal and distal histidine side-chains. The nicotinate and acetate ligands have been removed for clarity.

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

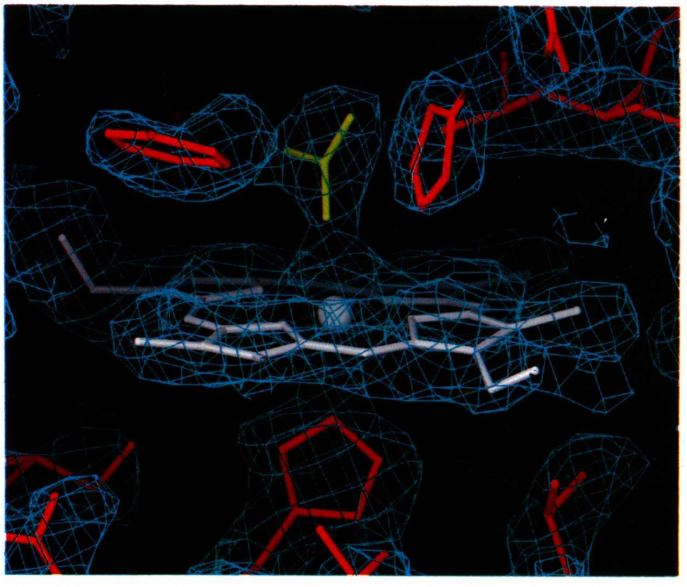


Figure 3. Electron density of the heme pocket in wild-type metLb*a*-acetate. The heme prosthetic group is shown in white, the acetate ligand in yellow and the protein in red. The distal histidine residue is within hydrogen-bonding distance of the carbonyl oxygen atom of the bound acetate group.

Lb*a* produces a large increase in k_{O_2}, whereas in myoglobin all of the mutations cause at least 100-fold increases in k_{O_2}. These results clearly show that the hydrogen bond between oxygen and His-E7 in leghemoglobin is very weak at neutral pH and not the main determinant of the high affinity of this protein for O_2, in agreement with the EPR results of Ikeda-Saito *et al.* (1981) and Lee *et al.* (1993). The three- to fourfold increases in k_{O_2} observed for the Val(E7) and Leu(E7) Lb*a* mutants are balanced by roughly equal increases in k_{O_2} so there is little net change in overall oxygen affinity. The Ala(E7) and Phe(E7) mutations produce an increase and decrease, respectively, in K_{O_2} with respect to the wild-type protein. These latter changes seem to imply that the E7 residue may sterically hinder the bound ligand. However, an exact interpretation of these results will require high-resolution crystal structures of the proteins. Regardless, it is clear that the polarity of His(E7) plays a much smaller role in regulating ligand binding to leghemoglobin than to myoglobin.

The effects of pH on oxygen binding to native, wild-type and E7 mutants of soybean leghemoglobin are presented in Table 3. As shown by Appleby *et al.* (1983), when the pH is decreased from 7 to 5, the rate constant for oxygen dissociation from Lb*a* decreases twofold with no change in the association rate constant. The resultant increase in affinity at low pH has been attributed to a strengthening of the hydrogen bond between His61 and bound ligand. The apolar mutants of Lb*a* show no pH dependence (Table 3). These mutagenesis results support the view that His61 plays a role in the pH dependence of O_2 binding to native leghemoglobin and that hydrogen bonding to the bound ligand is enhanced under acidic conditions.

CO and NO binding

The rate and equilibrium constants for carbon monoxide (CO) and nitric oxide (NO) binding to Lb*a*, sperm whale myoglobin and the correspond-

Table 2. The effects of apolar substitutions at position E7 on the rate and equilibrium constants of O_2 binding to soybean Lb*a* and sperm whale Mb at pH 7, 20°C

	k'_{O_2} $\mu M^{-1} s^{-1}$		k_{O_2} s^{-1}		K_{O_2} μM^{-1}	
Residue at 61(E7)	Lb	Mb	Lb	Mb	Lb	Mb
His(wild-type)	130	17	5.6	15	23	1.1
Ala	290	53	3.1	2300	94	0.023
Val	280	110	27	10000	10	0.011
Leu	400	98	24	4100	17	0.023
Phe	130	75	280	10000	0.45	0.0075

Data for sperm whale myoglobin were taken from Springer *et al.* (1994).

Motif Exhibit 1059, 622 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Table 3. Dependence of O_2 binding parameters on pH for native, wild-type, and Leu-E7 soybean Lb*a* at 20°C.

Soybean Lb	pH 7			pH 5		
	k'_{O_2} ($\mu M^{-1} s^{-1}$)	k_{O_2} (s^{-1})	K_{O_2} (μM^{-1})	k'_{O_2} ($\mu M^{-1} s^{-1}$)	k_{O_2} (s^{-1})	K_{O_2} (μM^{-1})
Native	130	5.6	23	140	2.7	51
Wild-type	130	5.6	23	130	2.2	60
H61A	290	3.1	94	250	2.7	92
H61V	280	27	10	280	28	10
H61L	400	24	17	470	25	19
H61F	130	280	0.45	120	230	0.54

ing Leu(E7) mutants are presented in Table 4. Replacement of the distal histidine residue with Leu increases the association rate constants for CO and NO binding to both proteins, but again the effects are much greater in myoglobin. The rate constants for CO and NO dissociation are largely unaffected by the Leu(E7) replacement. For both ligands, wild-type Lb*a* has larger association rate constants and smaller dissociation rate constants than wild-type myoglobin. These results point out that with or without a distal histidine residue, leghemoglobin is inherently much more reactive than myoglobin.

Autooxidation and hemin loss

A comparison of resistances to auto-oxidation and hemin dissociation is given in Table 5. Brantley *et al.* (1993) have shown that the rate constants for auto-oxidation of myoglobin mutants correlate inversely with O_2 affinity. Thus, we expected Lb*a* to auto-oxidize slowly due to its high O_2 affinity. However, the opposite effect was observed; k_{autoox} for Lb*a* is about fourfold greater than that of wild-type sperm whale Mb. This result does correlate with a much weaker hydrogen bond between bound O_2 and His(E7) in leghemoglobin. Brantley *et al.* (1993) showed that this polar interaction inhibits protonation of the Fe-O_2 complex and its subsequent dismutation into ferric iron and the neutral superoxide radical. Both myoglobin and leghemoglobin show a marked increase in k_{autoox} with decreasing pH, presumably due to protonation of bound O_2.

The rate constant for hemin dissociation from native and wild-type Lb*a* is ~6 h^{-1} at pH 7.0, 37°C. This rate is 600 times faster than that for myoglobin under the same conditions and, in combination with the increased auto-oxidation rate, results in a much less stable holoprotein (Hargrove & Olson, 1996). Lb*a* also has a much larger temperature dependence for hemin dissociation than myoglobin ($E_a \approx 33$ kcal/mol for aquometLb*a* *versus* $E_a \approx 8$ kcal/mol for sperm whale aquometMb at pH 7; Hargrove *et al.*, 1994; and unpublished results). As shown in Table 5, k_{-H} for met Lb*a* is ~0.8 h^{-1} at 20°C, which is roughly tenfold less than the rate observed at 37°C. The rate constant for hemin loss from myoglobin decreases by only a factor of ~2 when the temperature is decreased from 37 to 20°C at both pH 5.0 and pH 7.0 (Hargrove *et al.*, 1994; and unpublished results).

The rates of hemin loss from the distal histidine mutants of Lb*a* are too fast to be measured conveniently at 37°C. As a result, comparisons between mutant and wild-type leghemoglobins were measured at 20°C. The distal histidine residue of Lb*a* clearly stabilizes bound hemin, as shown by the H61L substitution, which increases the rate constant for hemin dissociation ~50-fold. A similar effect occurs when His(E7) is replaced by Leu in myoglobin. A large part of this effect is probably due to loss of the coordinated water molecule from the ferric H61L protein. Both Leu-E7 metMb and metLb*a* show a broad absorbance band centered at 395 nm, which is characteristic of 5-coordinated ferric proteins (Table 1; Quillin *et al.*, 1993). The pH dependence of hemin dissociation from Lb*a* is

Table 4. Comparison of the CO and NO binding parameters of soybean Lb*a* and sperm whale Mb at pH 7, 20°C

Protein	k'_{CO} ($\mu M^{-1} s^{-1}$)	k_{CO} (s^{-1})	K_{CO} (μM^{-1})	k'_{NO} ($\mu M^{-1} s^{-1}$)	k_{NO} (s^{-1})	K_{NO} (μM^{-1})
Native Lb	17	0.0078	2200	170	0.00002	9,000,000
Wild-type Lb	15	0.0084	1800	190	0.00002	9,000,000
H61L Lb	170	0.0024	72,000	320	0.00002	16,000,000
Wild-type Mb	0.51	0.019	27	22	0.00010	220,000
H64L Mb	26	0.024	1100	190	0.00013	1,500,000

Data for sperm whale myoglobin were taken from Springer *et al.* (1994).

Motif Exhibit 1059, 623 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Table 5. Comparison of the rate constants for hemin dissociation (k_{-H}) from and auto-oxidation (k_{autoox}) of soybean Lb*a* and sperm whale Mb

Protein	k_{autoox} (h^{-1}) at 37°C		k_{-H} (h^{-1})[b] at 20°C	
	pH 7	pH 5	pH 7	pH 5
Native Lb	0.2	1.6	0.7	1.3
Wild-type Lb	0.2	3.0	0.9	3.0
H61L Lb	0.3	6.6	47	45
Wild-type Mb	0.055	2.4	0.005[c](0.01)	0.4(1.0)
H64L Mb	10	200[a]	0.094[c](0.20)	5.2[c](11)

The k_{autoox} values for MbO_2 were taken from Brantley *et al.* (1993) and the values of k_{-H} for metMb from Hargrove *et al.* (1996b). Rates of auto-oxidation were measured in the presence of 1mM EDTA, catalase, and superoxide dismutase as described by Brantley *et al.* (1993). The k_{-H} values in parentheses for metMb were measured at 37°C by Hargrove *et al.* (1996b) and Hargrove *et al.* (1994).

[a] The value of k_{autoox} for Leu(E7) Mb at pH 5 was estimated assuming the same pH dependence as is observed for the Val(E7) Mb mutant by Brantley *et al.* (1993).

[b] The rates of hemin loss from leghemoglobin were measured at 20°C to allow comparison with mutant Lbs, which lose heme too rapidly at 37°C.

[c] The rates of hemin loss at 20°C were computed using the k_{-H} value at 37°C and an activation energy of 8.0 kcal/mol. The values in parentheses were measured at 37°C (Hargrove *et al.*, 1996b).

much smaller than that of myoglobin. The resistances of both proteins to auto-oxidation and hemin loss are similar at pH 5, whereas at pH 7 sperm whale myoglobin is much more stable.

Comparison with lupin LbII

As shown in Table 6, the association and dissociation rate constants for O_2 binding to lupin LbII are three- to fivefold greater than the corresponding parameters for soybean Lb*a*. This implies more rapid rates of ligand entry into and exit from the distal pocket of the lupin protein. Both proteins contain His(E7) and Phe(CD1). In lupin LbII the B10 and E11 residues are smaller, Phe and Val, respectively, *versus* Tyr and Leu in soybean leghemoglobin. However, this decrease in size does not appear to account for the ligand binding differences. Mutation of Leu65(E11) to Val has little effect on O_2 binding to soybean Lb*a*, and the Tyr30(B10) to Phe substitution causes decreases, not increases, in k'_{O_2} and k_{O_2} (Table 6). As suggested by Gibson *et al.* (1989), the differences in kinetic behavior among leghemoglobins appear to be due to amino acid changes remote from the ligand binding site. The increase in affinity due to the Y30F substitution is striking and suggests that an in-depth study of the role of this residue in ligand binding will be rewarding.

Table 6. Comparison of the rate constants for O_2 binding to soybean Lb*a* and Lupin LbII at pH 7, 20°C

Protein	k'_{O_2} ($\mu M^{-1}s^{-1}$)	k_{O_2} (s^{-1})	K_{O_2} (μM^{-1})
Lupin LbII	320	25	13
Soybean Lb*a*	130	5.6	23
Soybean Lb*a* (L65V)	180	5.8	30
Soybean Lb*a* (Y30F)	79	0.75	100

The rate constants for lupin LbII were taken from Gibson *et al.* (1989).

Discussion

The crystal structure of recombinant Lb*a* and the data reported in Tables 1 and 4 show that the protein expressed in *E. coli* TB1 cells is functionally and structurally identical with that of native Lb*a*. It is unclear why the N-terminal methionine residue is removed from this protein when wild-type myoglobin grown under the same conditions is expressed with the N-Met still attached (Springer & Sligar, 1987; Phillips *et al.*, 1990). The N-terminal amino acid sequences of the two proteins are similar (Lb*a*, VAFTE-; Mb, VLSEG-). There is no clear explanation for the difference in post-translational processing during constitutive expression.

Leghemoglobin is much less stable than myoglobin due to a fourfold higher rate of auto-oxidation and a ~600-fold higher rate of hemin loss at pH 7, 37°C. The rate constant for hemin loss of the H61L Lb*a* mutant, ~50 h^{-1}, is in the range of that for H93G sperm whale myoglobin, which lacks a covalent bond between the heme and globin (Hargrove *et al.*, 1996a). The increase in k_{-H} caused by the H61L mutation suggests that the distal histidine residue of Lb stabilizes bound heme by hydrogen bonding to a coordinated water molecule as is observed in myoglobin. The decreased stability of Lb*a* probably accounts for the lower yields in the constitutive expression (TB1) system relative to that of myoglobin. Purification of leghemoglobin in the presence of CO or other distal ligands prevents heme loss and greatly increases protein yields in both the constitutive and inducible expression systems.

The oxygen association rate constants for leghemoglobins are very fast, of the order of 10^8 M^{-1} s^{-1}, and are close to the rate expected for a diffusion-limited association reaction with an unhindered

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

model heme ($k'_{diffusion} \approx 5 \times 10^8\ M^{-1}\ s^{-1}$, Reisberg & Olson, 1980). Gibson *et al.* (1989) also suggested that the Lb rates may be the fastest possible for ligand association to a myoglobin-like protein. The high rates of ligand binding to Lb result in part from a very reactive heme iron atom as has been demonstrated with laser photolysis experiments (Stetzkowski *et al.*, 1985; Gibson *et al.*, 1989). Harutyunyan *et al.* (1995) have shown that the proximal histidine residue in lupin Lb is much more free to rotate and move toward the heme plane than in myoglobin. These results provide a structural explanation for the high reactivity of leghemoglobins. The heme pocket of Lb*a* is also more flexible and kinetically accessible than those of myoglobin and other heme proteins (Rohlfs *et al.*, 1988). This has been demonstrated by many techniques and is exemplified by the ability of Lb*a* to bind large ligands such as imidazole, nicotinate, and long-chain alkyl isocyanides, which do not bind tightly to myoglobin or hemoglobin (Stetzkowski *et al.*, 1979; Kong *et al.*, 1983; Mims *et al.*, 1983).

Even though the bimolecular rate constants are large, His(E7) still appears to offer some resistance to ligand association. The distal heme pocket of the deoxy forms of myoglobin and lupin leghemoglobin both contain a non-coordinated water molecule that hydrogen bonds to His(E7) (Quillin *et al.*, 1993; Harutyunyan *et al.*, 1995). This water molecule must be displaced before other ligands can bind. The increased association rate constants for oxygen binding to the H61A, H61V, and H61L mutants of Lb*a* are probably due to the loss of water in the deoxy forms of these proteins. The same mutations produce even larger increases in k'_{O_2} when constructed in sperm whale myoglobin (Quillin *et al.*, 1993; Table 2).

The functional evidence for a hydrogen bond between bound O_2 and His(E7) in leghemoglobin is equivocal at pH 7.0. The Ala, Val and Leu(E7) substitutions do not produce large decreases in K_{O_2} as is observed in myoglobin. However, the pH dependence of O_2 dissociation from leghemoglobin does appear to involve the distal histidine residue. Decreasing the pH from 7 to 5 causes a twofold increase in O_2 affinity when the E7 residue is histidine but has no effect when the E7 residue is apolar. This result suggests that hydrogen bonding occurs under acid conditions, but even at low pH, the strength of this interaction is still much less than that observed in myoglobin.

The β subunit of R-state human hemoglobin shows high rates of ligand association, and replacement of the distal histidine residue with either Gly or Phe has little effect on O_2 binding (Mathews *et al.*, 1989). Thus both soybean Lb*a* and R-state human β subunits have evolved mechanisms for stabilizing bound O_2 and inhibiting auto-oxidation that do not involve a strong hydrogen bond with His(E7). In contrast, strong favorable polar interactions occur in the distal pockets of all mammalian myoglobins and the α subunit of human hemoglobin. However, the distal histidine residue is very important for retaining heme in leghemoglobin and human β subunits. Replacing the distal histidine residue with smaller apolar residues causes ⩾50-fold increases in the rate of hemin loss from met-Lb*a* and ferric β subunits, explaining the high degree of conservation of His(E7) in these proteins (Table 5; Hargrove *et al.*, 1994).

Materials and Methods

A full-length cDNA for Lb*a* was isolated from a soybean nodule cDNA library using anti-Lb*a* antibodies as a probe (Ji *et al.*, 1994a; Sambrook *et al.*, 1989). Phage (λ-GT11) containing the soybean Lb*a* insert were placed in *E. coli*, harvested, and the DNA isolated. Lb*a* cDNA was excised and ligated into pBluescript K+ (Invitrogen). To obtain the coding region, polymerase chain reaction (PCR) primers containing a *Nco*I site at the starting ATG codon and an *Eco*RI restriction site downstream of the TAA stop codon were used to amplify the Lb*a* gene from the pBluescript-Lb*a* vector. PCR-amplified products were purified by agarose gel electrophoresis and ligated into a pCRII vector (Invitrogen). After transformation, an *E. coli* colony containing the pCRII-Lb*a* construct was used to obtain purified plasmid with the *Nco*I-Lb*a*-*Eco*RI sequence. This gene fragment was removed from pCRII, purified from an agarose gel, and ligated into pACYC184. A duplex oligonucleotide of the constitutive *lac* promoter used by Springer & Sligar (1987) was constructed with the following sequence of restriction sites: 5′ *Bsu*36I-*Kpn*I-*lac* promoter-*Nco*I 3′. The pACYC184-Lb*a* plasmid was cut with *Nco*I and *Bsu*36I, and the oligonucleotide containing the *lac* promoter was ligated to the Lb*a* cDNA. This *lac* promoter-Lb*a* construct was removed from pACYC184 and ligated into pEMBL19+ using the flanking *Kpn*I/*Eco*RI sites.

The initial cDNA product cloned in this manner contained two base-pair differences compared to the Lb*a* sequence reported by Hyldig-Nielson *et al.* (1982). These changes were probably introduced during the PCR reaction used to amplify the cDNA. The gene was corrected to the native sequence by two rounds of site-directed mutagenesis using the Kunkel (1985) method. The resulting plasmid, pLb*a*, contains the native Lb*a* cDNA sequence in pEMBL19+ under control of the *lac* promoter (Figure 1). The Kunkel method of mutagenesis was also used to generate the B10, E7 and E11 mutants starting from the "corrected" wild-type Lb*a* gene.

Expression of wild-type and mutant Lb*a*s

E. coli TB-1 cells were transformed with pLb*a* and grown in LB medium (with ampicillin) at 37°C. The final yield of purified leghemoglobin (~0.6 mg/l) was roughly one-seventh of that for sperm whale myoglobin expressed under the same conditions (~4 mg/l). H61F, H61V and H61A proteins showed no appreciable expression in the *lac* promoter-TB1 system. Consequently, the *Nco*I-Lb*a*-*Eco*RI fragment was cloned into pET28a (Novagen) for high-level, inducible expression in *E. coli* BL21DE3 cells. Shake flasks of BL21DE3 cells transformed with the pET28a-Lb*a* vector were grown to an absorbance of ~0.6 in 2 × YT medium at 33°C, induced with 1 mM isopropyl-β,D-thiogalactopyramoside and incubated for an additional four hours. Apo-Lb*a* was found in the supernatant and pellet of the cell lysis mixture. Protein was re-

Motif Exhibit 1059, 625 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

covered from the lysis pellet by dissolution in 6 M guanidinium chloride followed by dialysis into 20 mM Tris (pH 8.0). Holoprotein was reconstituted by the addition of CO-heme or imidazole-hemin to the soluble and pellet fractions of the lysis mixture. The final yield of wild-type or mutant Lb*a* was ~10 mg/l of medium. Reconstituted holo-Lb*a* was precipitated with 90% saturated ammonium sulfate for further purification.

Regardless of the method of expression, recombinant Lb*a* was purified as described by Jun *et al.* (1994) with the following modifications. The 90% saturated ammonium sulfate Lb*a* pellet was resuspended in 1.8 M ammonium sulfate and loaded directly onto a phenyl Sepharose column, and the protein was not oxidized prior to G-75 chromatography. Purified Lb*a* was in the CO or met-imidazole form (depending on the method of reconstitution) and had a Soret to 280 nm absorbance ratio greater than 4. Native Lb*a* was purified from soybean root nodules using the method of Jun *et al.* (1994). N-terminal sequencing of the purified Lbs was performed at the UNL-Protein Core Facility on an ABI-494 Procise sequencer using a manufacturer-supplied protocol.

Crystallography of wild-type Lb*a*

Crystals of the acetate complex of wild-type metLb*a* were grown at room temperature to 0.5 mm × 0.05 mm × 0.05 mm in three to five days using the hanging drop method. The starting concentration of metLb*a* was 1 mM in 0.1 M sodium phosphate (pH 6.0). The precipitant was 2.4 M ammonium sulfate, 0.5 M ammonium acetate at pH 6.8 (Vainshtein *et al.*, 1975). X-ray diffraction data were collected at room temperature on an R-axis IIc imaging plate system using copper Kα radiation from a Siemens rotating anode operating at 50 mV and 90 mA. Crystals were in the $P2_12_12_1$ space group. The data were 90% complete at the limiting resolution of 2.2 Å, with an R_{merge} of 10.3%. Phases were determined by molecular replacement using XPLOR (Brünger, 1987) starting with the coordinates for the structure of Lb*a*-nicotinate provided by Dr Paul Ellis (PDB file 1fsl; Ollis *et al.*, 1983; Ellis *et al.*, 1997). Two molecules related by a 2-fold non-crystallographic symmetry axis were present in the asymmetric unit. After several cycles of refinement, manual fitting and solvent addition, the crystallographic *R*-factor converged to 19.8% (R_{free} = 29.7%) with root-mean-square deviations of 0.018 Å in bond lengths, 1.93° in bond angles, 21.36° in dihedral angles and 2.29° in improper angles. Molecule a of the two molecules in the asymmetric unit was used to produce Figures 2 and 3. The coordinates of the structure of wild-type metLb*a*-acetate have been assigned number 1bin in the Protein Data Bank at the Brookhaven National Laboratory.

Spectroscopic and kinetic characterization of wild-type and mutant Lb*a*

Absorbance spectra were measured using a SLM 3000 diode array spectrophotometer. LbCO was prepared by diluting metLb into a cuvette containing 20 mM Tris (pH 8.0) buffer equilibrated with 1 atm CO and an excess of sodium dithionite. MetLb was formed by the reaction of ferrous Lb with potassium ferricyanide followed by chromatography on a Sephadex G-25 column equilibrated in 20 mM Tris (pH 8.0). LbO_2 was produced by the reduction of metLb with sodium dithionite followed by G-25 chromatography in the presence of oxygen. DeoxyLb*a* was produced by the addition of several granules of sodium dithionite to a cuvette containing metLb in anaerobic buffer.

The rate constants for O_2, CO and NO binding were measured using established rapid mixing and laser photolysis methods (Olson, 1981; Quillin *et al.*, 1995; Moore & Gibson, 1976). O_2 and CO association rate constants at pH 5.0 were measured in 100 mM sodium acetate (pH 5.0). O_2 dissociation at pH 5 was measured by mixing a LbO_2 sample in 10 mM potassium phosphate (pH 7.0) with 100 mM sodium acetate (pH 5.0), equilibrated with 1 atm CO (Appleby *et al.*, 1983).

Rate constants for hemin dissociation and auto-oxidation were measured as described by Hargrove *et al.* (1994) and Brantley *et al.* (1993), respectively. Hemin dissociation from H61L Lb*a* was measured in a Gibson-Dionex stopped-flow apparatus as described by Hargrove *et al.* (1996a). Mes (2-(*N*-morpholino)ethanesulfonic acid) was used as the buffer at pH 5 for measurement of hemin dissociation because acetate coordination to metLb markedly inhibits this reaction.

Acknowledgments

This work was supported by United States Public Health Service grants GM35649 and HL-47020, grant C-612 from the Robert A. Welch Foundation, and the W. M. Keck Foundation (J.S.O.), grant 003604-025 from the State of Texas Advanced Technology Program (G.N.P./J.S.O.), a National Institutes of Health Postdoctoral Fellowship ARO8355 (E.A.B.), NSF grant OSR-92552255 (R.V.K.), USDA-CREES 95-37305-2441 (R.V.K./G.S.), and the Nebraska Research Initiative (G.S.). R.A.P. was partially supported by DGAPA (Direccion General de Asuntos del Personal Academico)-UNAM and CoNaCyT (Consejo Nacional de Cieneia y Tecnologia), Mexico. We also thank M. Ikeda-Saito, C. Appleby and A. Wilkinson for reading the original manuscript and making helpful suggestions.

References

Appleby, C. A. (1984). Leghemoglobins and *Rhizobium* respiration. *Annu. Rev. Plant Physiol.* **35**, 443–478.

Appleby, C. A. (1992). The origin and functions of haemoglobin in plants. *Sci. Prog.* **76**, 365–398.

Appleby, C. A., Bradbury, J. H., Morris, R. J., Wittenberg, B. A., Wittenberg, J. B. & Wright, P. E. (1983). Leghemoglobin. *J. Biol. Chem.* **258**, 2254–2259.

Brantley, R. E., Smerdon, S. J., Wilkinson, A. J., Singleton, E. W. & Olson, J. S. (1993). The mechanism of autooxidation of myoglobin. *J. Biol. Chem.* **268**, 6995–7010.

Brünger, A. (1987). Crystallographic *R* factor refinement by molecular dynamics. *Science*, **235**, 458–460.

Ellis, P. J., Appleby, C. A., Guss, J. M., Hunter, W. N., Ollis, D. L. & Freeman, H. C. (1997). The crystal structure of ferric soybean leghemoglobin *a* nicotinate at 2.3 Å resolution. *Acta. Crystallog. sect. D*, in the press.

Fuchsman, W. H. & Appleby, C. A. (1979). CO and O_2 complexes of soybean leghemoglobins: pH effects upon infrared and visible spectra. Comparisons with CO and O_2 complexes of myoglobin and hemoblobin. *Biochemistry*, **18**, 1309–1321.

Motif Exhibit 1059, 626 of 653

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Gibson, Q. H., Wittenberg, J. B., Wittenberg, B. A., Bogusz, D. & Appleby, C. A. (1989). The kinetics of ligand binding to plant hemoglobins. *J. Biol. Chem.* **264**, 100–107.

Hargrove, M. S. & Olson, J. S. (1996). The stability of holomyoglobin is determined by heme affinity. *Biochemistry*, **35**, 11310–11318.

Hargrove, M. S., Singleton, E. W., Quillin, M. L., Ortiz, L. A., Phillips, G. N., Jr, Mathews, A. J. & Olson, J. S. (1994). His(E7)Tyr as a reagent for measurement of hemin loss. *J. Biol. Chem.* **269**, 4207–4214.

Hargrove, M. S., Barrick, D. & Olson, J. S. (1996a). The association rate constant for heme binding to globin is independent of protein structure. *Biochemistry*, **35**, 11293–11299.

Hargrove, M. S., Wilkinson, A. J. & Olson, J. S. (1996b). Structural factors governing hemin dissociation from metmyoglobin. *Biochemistry*, **35**, 11300–11309.

Harutyunyan, E. H., Safonova, T. N., Kuranova, I. P., Popov, A. N., Teplyakov, A. V., Obmolova, G. V., Rusakov, A. A., Vainshtein, B. K., Dodson, G. G., Wilson, J. C. & Perutz, M. F. (1995). The structure of deoxy- and oxy-leghaemoglobin from lupin. *J. Mol. Biol.* **251**, 104–115.

Hyldig-Nielsen, J. J., Jensen, E. O., Paludan, K., Wiborg, O., Garret, R., Jorgensen, P. & Marcker, K. A. (1982). The primary structures of two leghemoglobin genes from soybean. *Nucl. Acids Res.* **10**, 689–695.

Ikeda-Saito, M., Hori, H., Inubushi, T. & Yonetani, T. (1981). Studies on cobalt myoglobins and hemoglobins. *J. Biol. Chem.* **256**, 10267–10271.

Ji, L., Becana, M., Sarath, G. & Klucas, R. V. (1994a). Cloning and sequence analysis of a cDNA encoding ferric leghemoglobin reductase from soybean nodules. *Plant Physiol.* **104**, 453–459.

Ji, L., Becana, M., Sarath, G., Shearman, L. & Klucas, R. V. (1994b). Over-expression in *Escherichia coli* and characterization of a soybean ferric leghemoglobin reductase. *Plant Physiol.* **106**, 203–209.

Jun, H.-K., Sarath, G. & Wagner, F. W. (1994). Detection and purification of modified leghemoglobins from soybean root nodules. *Plant Sci.* **100**, 31–40.

Kloek, A. P., Yang, J., Mathews, F. S., Frieden, C. & Goldberg, D. E. (1993). The tyrosine B10 hydroxyl is crucial for oxygen avidity of *Ascaris* hemoglobin. *J. Biol. Chem.* **269**, 2377–2379.

Kong, S. B., Cutnell, J. D. & La Mar, G. N. (1983). Proton nuclear magnetic resonance study of the dynamic stability of the heme pocket of soybean leghemoglobin *a*. *J. Biol. Chem.* **258**, 3843–3849.

Kunkel, T. A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl Acad. Sci. USA*, **82**, 488–492.

Lee, H. C., Wittenberg, J. B. & Peisach, J. (1993). Role of hydrogen bonding to bound dioxygen in soybean leghemoglobin. *Biochemistry*, **32**, 11500–11506.

Martin, K. D., Saari, L., Guang-Xin, W., Wang, T., Parkhurst, L. J. & Klucas, R. V. (1990). Kinetics and thermodynamics of oxygen, CO, and azide binding by the subcomponents of soybean leghemoglobin. *J. Biol. Chem.* **265**, 19588–19593.

Mathews, A. J., Rohlfs, R. J., Olson, J. S., Tame, J., Renaud, J. P. & Nagai, K. (1989). The effects of E7 and E11 mutations on the kinetics of ligand binding to R state human hemoglobin. *J. Biol. Chem.* **264**, 16573–16583.

Mathews, A. J., Olson, J. S., Renaud, J.-P., Tame, J. & Nagai, K. (1991). The assignment of carbon monoxide association rate constants to the a and β subunits in native and mutant human deoxyhemoglobin tetramers. *J. Biol. Chem.* **266**, 21631–21693.

Mims, M. P., Porras, A. G., Olson, J. S., Noble, R. W. & Peterson, J. A. (1983). Ligand binding to heme proteins. *J. Biol. Chem.* **258**, 14219–14232.

Moore, E. G. & Gibson, Q. H. (1976). Cooperativity in the dissociation of nitric oxide from hemoglobin. *J. Biol. Chem.* **251**, 2788–2794.

Nagai, K., Luisi, B., Shih, D., Miyazaki, G., Imai, K., Poyart, C., De Young, A., Kwiatkowsky, L., Noble, R. W., Lin, S.-H. & Yu, N.-T. (1987). Distal residues in the oxygen binding site of haemoglobin studied by protein engineering. *Nature*, **329**, 858–860.

Ollis, D. L., Appleby, C. A., Colman, P. M., Cutten, A. E., Guss, J. M., Venkatappa, M. P. & Freeman, H. C. (1983). Crystal structure of soybean ferric leghaemoglobin *a* nicotinate at a resolution of 3.3 Å. *Aust. J. Chem.* **36**, 451–468.

Olson, J. S. (1981). Stopped-flow, rapid mixing measurements of ligand binding to hemoglobin and red cells. *Methods Enzymol.* **76**, 631–652.

Olson, J. S. & Phillips, G. N., Jr (1996). Kinetic pathways and barriers for ligand binding to myoglobin. *J. Biol. Chem.* **271**, 17593–17596.

Phillips, G. N., Jr, Arduini, R. M., Springer, B. A. & Sligar, S. G. (1990). Crystal structure of myoglobin from a synthetic gene. *Proteins: Struct. Funct. Genet.* **7**, 358–365.

Quillin, M. L., Arduini, R. M., Olson, J. S. & Phillips, G. N., Jr (1993). High-resolution crystal structures of distal histidine mutants of sperm whale myoglobin. *J. Mol. Biol.* **234**, 140–155.

Quillin, M. L., Li, T., Olson, J. S., Phillips, G. N., Jr, Dou, Y., Ikeda-Saito, M., Regan, R., Carlson, M., Gibson, Q. H., Li, H. & Elber, R. (1995). Structural and functional effects of apolar mutations of the distal valine in myoglobin. *J. Mol. Biol.* **245**, 416–436.

Reisberg, P. I. & Olson, J. S. (1980). Rates of isonitrile binding to the isolated α and β subunits of human hemoglobin. *J. Biol. Chem.* **255**, 4151–4158.

Rohlfs, R. J., Olson, J. S. & Gibson, Q. H. (1988). A comparison of the geminate recombination kinetics of several monomeric heme proteins. *J. Biol. Chem.* **263**, 100–107.

Saari, L. L. & Klucas, R. V. (1984). Ferric leghemoglobin reductase from soybean root nodules. *Arch. Biochem. Biophys.* **231**, 102–113.

Sambrook, J, Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edit, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Springer, B. A. & Sligar, S. G. (1987). High-level expression of sperm whale myoglobin in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **84**, 8961–8965.

Springer, B. A., Egeberg, K. D., Sligar, S. G., Rohlfs, R. J., Mathews, A. J. & Olson, J. S. (1989). Discrimination between oxygen and carbon monoxide and inhibition of autooxidation by myoglobin. *J. Biol. Chem.* **264**, 3057–3060.

Springer, B. A., Sligar, S. G., Olson, J. S. & Phillips, G. N., Jr (1994). Mechanisms of ligand recognition in myoglobin. *Chem. Rev.* **94**, 699–714.

Stetzkowski, F., Cassoly, R. & Banerjee, R. (1979). Binding of alkylisocyanides with soybean leghemoglobin. *J. Biol. Chem.* **254**, 11351–11356.

Stetzkowski, F., Banerjee, R., Marden, M. C., Beece, D. K., Bowne, S. F., Doster, W., Eisenstein, L., Frauenfelder, H., Reinisch, L., Shyamsunder, E. & Jung, C. (1985). Dynamics of dioxygen and carbon monoxide binding to soybean leghemoglobin. *J. Biol. Chem.* **260**, 8803–8809.

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Vainshtein, B. K., Arutyunyan, E. G., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavlovskii, A. G., Grebenko, A. I. & Konareva, N. V. (1975). X-ray study of leghemoglobin. I. Purification, crystallization, and preparation of derivatives containing heavy atoms. *Sov. Phys. Crystallog.* **19**, 598–601.

Edited by K. Nagai

(Received 9 September 1996; received in revised form 27 November 1996; accepted 4 December 1996)

Ex. 1054 Benson 2014

PRINT ISSN: 0305-1048
ONLINE ISSN: 1362-4962

Nucleic Acids Research

VOLUME 42 DATABASE ISSUE JANUARY 1, 2014

www.nar.oxfordjournals.org



Downloaded from https://academic.oup.com/nar/article/42/D1/i/1041193 by University of Wisconsin-Madison Libraries user on 09 January 2023

OXFORD
UNIVERSITY PRESS

Open Access
No barriers to access – all articles freely available online

MIX
Paper from responsible sources
FSC® C007207
www.fsc.org

Nucleic Acids Research

OXFORD UNIVERSITY PRESS

Downloaded from https://academic.oup.com/nar/article/42/D1/ii/1041196 by University of Wisconsin-Madison Libraries user on 09 January 2023

Executive Editors

Senior Editors: K.R.Fox, *Southampton, UK*; foxnar@soton.ac.uk B.Stoddard, *Seattle, WA, USA*; stoddnar@fhcrc.org

M.Madan Babu, *Cambridge, UK*
narlmb@mrc-lmb.cam.ac.uk

D.Corey, *Dallas, TX, USA*
david.corey@utsouthwestern.edu

M.Wegner, *Erlangen, Germany*
m.wegner@biochem.uni-erlangen.de

J.M.Bujnicki, *Warsaw, Poland*
iamb@genesilico.pl

W.Dynan, *Atlanta, GA, USA*
wsd.nar@gmail.com

E.Westhof, *Strasbourg, France*
nar@ibmc.u-strasbg.fr

M.Churchill, *Aurora, CO, USA*
nar@uchsc.edu

A.D.Sharrocks, *Manchester, UK*
nar@manchester.ac.uk

J.A.Wise, *Cleveland, OH, USA*
jaw17@po.cwru.edu

Methods Editors: G.Sczakiel, *Lübeck, Germany*; sczakiel@imm.uni-luebeck.de A.R.Kimmel, *Bethesda, MD, USA*; ark.nar@gmail.com
Database Issue Editor: M.Galperin, *Bethesda, MD, USA*; nardatabase@gmail.com
Web Server Issue Editor: G.Benson, *Boston, MA, USA*; gbenson@bu.edu

Editorial Board

F.Allain, *Zurich, Switzerland*
C.Bailly, *Lille, France*
A. Bateman, *Hinxton, UK*
P.B.Becker, *Munich, Germany*
M.Belfort, *Albany, NY, USA*
B.Berkhout, *Amsterdam, The Netherlands*
E.H.Bresnick, *Madison, WI, USA*
J.F.Caceres, *Edinburgh, UK*
J.S.Carroll, *Cambridge, UK*
X.Cheng, *Atlanta, GA, USA*
J.J.Collins, *Boston, MA, USA*
J.Dinman, *College Park, MD, USA*
G.Divita, *Montpellier, France*
S.Eddy, *Ashburn, VA, USA*
M.Egli, *Nashville, TN, USA*
D.R.Engelke, *Ann Arbor, MI, USA*
A.Engelman, *Boston, MA, USA*
M.J.Gait, *Cambridge, UK*
H.U.Göringer, *Darmstadt, Germany*
B.Gottgens, *Cambridge, UK*
T.Grange, *Paris, France*
S.Griffiths-Jones, *Manchester, UK*
W.Gruissem, *Zurich, Switzerland*
J.G.Hacia, *Los Angeles, CA, USA*
P.Hsieh, *Bethesda, MD, USA*
P.A.Jeggo, *Brighton, UK*
A.Khvorova, *Lafayette, CO, USA*
R.Kierzek, *Poznan, Poland*
H.Klein, *New York, NY, USA*
D.Klostermeier, *Muenster, Germany*
E.T.Kool, *Stanford, CA, USA*
E.Koonin, *Bethesda, MD, USA*
A.N.Lane, *Louisville, KY, USA*
B.Lehner, *Barcelona, Spain*
B.Lenhard, *London, UK*
D.M.J.Lilley, *Dundee, UK*
S.Linn, *Berkeley, CA, USA*
L.J.Maher, *Rochester, MN, USA*
D.Mann, *Newcastle-upon-Tyne, UK*
A.Maxwell, *Norwich, UK*
S.Mueller, *Greifswald, Germany*
S.Neidle, *London, UK*
R.Parker, *Tucson, AZ, USA*
M.R.Parthun, *Columbus, OH, USA*
R.Pillai, *Grenoble, France*
M.Porteus, *Stanford, CA, USA*
N.Proudfoot, *Oxford, UK*
R.J.Roberts, *Ipswich, MA, USA*
D.B.Roth, *Philadelphia, PA, USA*
T.D.Schneider, *Frederick, MD, SA*
B.Schwer, *New York, NY, USA*
B.Séraphin, *Gif sur Yvette, France*
P.M.Sharp, *Edinburgh, UK*
S.K.Silverman, *Urbana, IL, USA*
H.Siomi, *Tokyo, Japan*
C.Smolke, *Stanford, CA, USA*
I.Stancheva, *Edinburgh, UK*
A.Stasiak, *Lausanne, Switzerland*
N.Sugimoto, *Kobe, Japan*
J.Vogel, *Würzburg, Germany*
M.C.Williams, *Boston, MA, USA*
S.H.Wilson, *Research Triangle Park, NC, USA*
R.D.Wood, *Smithville, TX, USA*
S.A.Woodson, *Baltimore, MD, USA*
S.Yokoyama, *Tokyo, Japan*

Founding Editors

R.T.Walker, *Birmingham, UK*
D.Söll, *New Haven, CT, USA*
A.S.Jones, *Birmingham, UK*

Editorial and Production

Editorial Manager
Martine Bernardes-Silva
nar@soton.ac.uk

Production Editor
Oliver Barham
narese@oup.com

Subscriptions

A subscription to *Nucleic Acids Research* comprises 22 issues. All prices include postage, and for subscribers outside the UK delivery is by Standard Air. *Nucleic Acids Research* operates a model of continuous online publication, more details of which can be found at www.nar.oxfordjournals.org

Annual Subscription Rate (Volume 42, 22 issues, 2014)
Institutional
Print only: £2584.00/$4929.00/€3699.00
Institutional membership online only: £2396.00/$4793.00/€3596.00
Personal
Print: £618.00/$1237.00/€929.00
Please note: US$ rate applies to US and Canada, Euros applies to Europe, UK£ applies to UK and Rest of World.

Online access was made completely free of charge from 2005. An institutional membership provides discounted publication charges for corresponding authors based at the member institution. The print subscription does not entitle authors to discounted publication charges. See www.nar.oxfordjournals.org/openaccess. There may be other subscription rates available, for a complete listing please visit www.oxfordjournals.org/nar/access_purchase/price_list.html.

Full prepayment, in the correct currency, is required for all orders. Orders are regarded as firm and payments are not refundable. Subscriptions are accepted and entered on a complete volume basis. Claims cannot be considered more than FOUR months after publication or date of order, whichever is later. All subscriptions in Canada are subject to GST. Subscriptions in the EU may be subject to European VAT. If registered, please supply details to avoid unnecessary charges. For subscriptions that include online versions, a proportion of the subscription price may be subject to UK VAT. Personal rate subscriptions are only available if payment is made by personal cheque or credit card and delivery is to a private address.

The current year and two previous years' issues are available from Oxford University Press. Previous volumes can be obtained from the Periodicals Service Company, 11 Main Street, Germantown, NY 12526, USA. Email: psc@periodicals.com. Tel: +1 (518) 537 4700. Fax: +1 (518) 537 5899. Back volumes of this journal are available in 16 mm microfilm, 35 mm microfilm and 105 microfiche from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, USA. Copies of articles published are also available from UMI.

For further information, please contact: Journals Customer Service Department, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK. Email: jnls.cust.serv@oup.com. Tel (and answerphone outside normal working hours): +44 (0)1865 353907. Fax: +44 (0)1865 353485. **In the US, please contact:** Journals Customer Service Department, Oxford University Press, 2001 Evans Road, Cary, NC 27513, USA. Email: jnlorders@oup.com. Tel (and answerphone outside normal working hours): 800 852 7323 (toll-free in USA/Canada). Fax: 919 677 1714. **In Japan, please contact:** Journals Customer Service, Oxford University Press, 4-5-10-8F Shiba, Minato-ku, Tokyo 108-8386, Japan. Tel: +81 3 5444 5858. Fax: +81 3 3454 2929. Email: custserv.jp@oup.com

Methods of payment. (i) Cheque (payable to Oxford University Press, to Oxford University Press, Cashiers Office, Great Clarendon Street, Oxford OX2 6DP, UK) in GB£ Sterling (drawn on a UK bank), US$ Dollars (drawn on a US bank), or EU'Euros. (ii) Bank transfer to Barclays Bank Plc, Oxford Group Office, Oxford (bank sort code 20-65-18) (UK), overseas only Swift code BARC GB 22 (GB£ Sterling to account no. 70299332, IBAN GB89BARC20651870299332; US$ Dollars to account no. 66014600, IBAN GB27BARC20651866014600; EU'Euros to account no. 78923655, IBAN GB16BARC20651878923655). (iii) Credit card (Mastercard, Visa, Switch or American Express).

Nucleic Acids Research is published twice monthly by Oxford University Press, Oxford, UK. Annual subscription price is £2584.00/$4929.00/€3699.00. *Nucleic Acids Research* is distributed by Mercury Media Processing, 1634 E. Elizabeth Ave, Linden, NJ 07036, USA. Periodicals postage paid at Rahway, NJ and at additional entry points. Issue Date: **January 28, 2014**.

US Postmaster: send address changes to *Nucleic Acids Research* (ISSN 0305-1048), c/o Mercury Media Processing, 1634 E. Elizabeth Ave, Linden, NJ 07036, USA.

For information about DOIs and to resolve them, please visit http://dx.doi.org/

Oxford Journals Environmental and Ethical Policies
Oxford Journals, a division of Oxford University Press, is committed to working with the global community to bring the highest quality research to the widest possible audience. Oxford Journals will protect the environment by implementing environmentally friendly policies and practices wherever possible. Please see http://www.oxfordjournals.org/ethicalpolicies.html for further information on environmental and ethical policies.

Permissions
For information on how to request permissions to reproduce articles/information from this journal, please visit http://www.oxfordjournals.org/jnls/permissions/.

Advertising
Advertising, inserts and artwork enquiries should be addressed to Advertising and Special Sales, Oxford Journals, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK. Tel: +44 (0)1865 354767; Email: jnlsadvertising@oup.com.

© Oxford University Press 2014

All rights reserved; 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 Publishers, or a licence permitting restricted copying issued in the UK by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1P 9HE, or in the USA by the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Typeset by Cenveo Publisher Services, Bangalore, India.
Printed by C.O.S. Printers Pte Ltd, Singapore.

Cover: Understanding the complex relationships between DNA sequence, its 3D structure, and protein-DNA binding specificity is at the frontier of research that integrates genomics with structural biology. Rohs and colleagues present TFBS shape, a new database (see Yang *et al*. *Nucleic Acids Res*. **42**, D148–D155) that disentangles these relationships. This task is like solving a Rubik's cube (top face: DNA sequences with transcription factor binding sites; left face: 3D structure of protein-DNA complex; front face: heat map representing minor groove width patterns selected by a transcription factor in a high-throughput experiment). The cover figure was designed by Tianyin Zhou and Lin Yang (University of Southern California) based on a photo by Konstantin Datz.

Downloaded from https://academic.oup.com/nar/article/42/D1/iii/1041200 by University of Wisconsin-Madison Libraries user on 09 January 2023

Nucleic Acids Research

Contents

Volume 42, Database issue, January 1, 2014

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

I. Nucleic acid sequence, structure and regulation

The 2014 *Nucleic Acids Research* Database Issue and an updated NAR online Molecular Biology Database Collection	X.M.Fernández-Suárez, D.J.Rigden and M.Y.Galperin	D1–D6
Database resources of the National Center for Biotechnology Information	NCBI Resource Coordinators, A.Acland, R.Agarwala, T.Barrett, J.Beck, D.A.Benson, C.Bollin, E.Bolton, S.H.Bryant, K.Canese, D.M.Church, K.Clark, M.DiCuccio, I.Dondoshansky, S.Federhen, M.Feolo, L.Y.Geer, V.Gorelenkov, M.Hoeppner, M.Johnson, C.Kelly, V.Khotomlianski, A.Kimchi, M.Kimelman, P.Kitts, S.Krasnov, A.Kuznetsov, D.Landsman, D.J.Lipman, Z.Lu, T.L.Madden, T.Madej, D.R.Maglott, A.Marchler-Bauer, I.Karsch-Mizrachi, T.Murphy, J.Ostell, C.O'Sullivan, A.Panchenko, L.Phan, D.P.K.D.Pruitt, W.Rubinstein, E.W.Sayers, V.Schneider, G.D.Schuler, E.Sequeira, S.T.Sherry, M.Shumway, K.Sirotkin, K.Siyan, D.Slotta, A.Soboleva, V.Soussov, G.Starchenko, T.A.Tatusova, B.W.Trawick, D.Vakatov, Y.Wang, M.Ward, W.John Wilbur, E.Yaschenko and K.Zbicz	D7–D17
The European Bioinformatics Institute's data resources 2014	C.Brooksbank, M.T.Bergman, R.Apweiler, E.Birney and J.Thornton	D18–D25
The genome portal of the Department of Energy Joint Genome Institute: 2014 updates	H.Nordberg, M.Cantor, S.Dusheyko, S.Hua, A.Poliakov, I.Shabalov, T.Smirnova, I.V.Grigoriev and I.Dubchak	D26–D31
GenBank	D.A.Benson, K.Clark, I.Karsch-Mizrachi, D.J.Lipman, J.Ostell and E.W.Sayers	D32–D37
Assembly information services in the European Nucleotide Archive	N.Pakseresht, B.Alako, C.Amid, A.Cerdeño-Tárraga, I.Cleland, R.Gibson, N.Goodgame, T.Gur, M.Jang, S.Kay, R.Leinonen, W.Li, X.Liu, R.Lopez, H.McWilliam, A.Oisel, S.Pallreddy, S.Plaister, R.Radhakrishnan, S.Rivière, M.Rossello, A.Senf, N.Silvester, D.Smirnov, S.Squizzato, P.t.Hoopen, A.L.Toribio, D.Vaughan, V.Zalunin and G.Cochrane	D38–D43
DDBJ progress report: a new submission system for leading to a correct annotation	T.Kosuge, J.Mashima, Y.Kodama, T.Fujisawa, E.Kaminuma, O.Ogasawara, K.Okubo, T.Takagi and Y.Nakamura	D44–D49
Updates to BioSamples database at European Bioinformatics Institute	A.Faulconbridge, T.Burdett, M.Brandizi, M.Gostev, R.Pereira, D.Vasant, U.Sarkans, A.Brazma and H.Parkinson	D50–D52

Continued

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes	S.Geisen, G.Barturen, Á.M.Alganza, M.Hackenberg and J.L.Oliver	D53–D59
	uORFdb—a comprehensive literature database on eukaryotic uORF biology	K.Wethmar, A.Barbosa-Silva, M.A.Andrade-Navarro and A.Leutz	D60–D67
	miRBase: annotating high confidence microRNAs using deep sequencing data	A.Kozomara and S.Griffiths-Jones	D68–D73
S	**miRNEST 2.0: a database of plant and animal microRNAs**	M.W.Szcześniak and I.Makałowska	D74–D77
S	**miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions**	S.-D.Hsu, Y.-T.Tseng, S.Shrestha, Y.-L.Lin, A.Khaleel, C.-H.Chou, C.-F.Chu, H.-Y.Huang, C.-M.Lin, S.-Y.Ho, T.-Y.Jian, F.-M.Lin, T.-H.Chang, S.-L.Weng, K.-W.Liao, I.-E.Liao, C.-C.Liu and H.-D.Huang	D78–D85
	PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways	A.Bhattacharya, J.D.Ziebarth and Y.Cui	D86–D91
S	**starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data**	J.-H.Li, S.Liu, H.Zhou, L.-H.Qu and J.-H.Yang	D92–D97
	NONCODEv4: exploring the world of long non-coding RNA genes	C.Xie, J.Yuan, H.Li, M.Li, G.Zhao, D.Bu, W.Zhu, W.Wu, R.Chen and Y.Zhao	D98–D103
S	**NPInter v2.0: an updated database of ncRNA interactions**	J.Yuan, W.Wu, C.Xie, G.Zhao, Y.Zhao and R.Chen	D104–D108
	RADAR: a rigorously annotated database of A-to-I RNA editing	G.Ramaswami and J.B.Li	D109–D113
	The Nucleic Acid Database: new features and capabilities	B.Coimbatore Narayanan, J.Westbrook, S.Ghosh, A.I.Petrov, B.Sweeney, C.L.Zirbel, N.B.Leontis and H.M.Berman	D114–D122
S	**RNA Bricks—a database of RNA 3D motifs and their interactions**	G.Chojnowski, T.Waleń and J.M.Bujnicki	D123–D131
S	**SMMRNA: a database of small molecule modulators of RNA**	A.Mehta, S.Sonam, I.Gouri, S.Loharch, D.K.Sharma and R.Parkesh	D132–D141
S	**JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles**	A.Mathelier, X.Zhao, A.W.Zhang, F.Parcy, R.Worsley-Hunt, D.J.Arenillas, S.Buchman, C.-y.Chen, A.Chou, H.Ienasescu, J.Lim, C.Shyr, G.Tan, M.Zhou, B.Lenhard, A.Sandelin and W.W.Wasserman	D142–D147
S	**TFBSshape: a motif database for DNA shape features of transcription factor binding sites**	L.Yang, T.Zhou, I.Dror, A.Mathelier, W.W.Wasserman, R.Gordân and R.Rohs	D148–D155
S	**CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria**	S.Kılıç, E.R.White, D.M.Sagitova, J.P.Cornish and I.Erill	D156–D160
	The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*	M.C.Teixeira, P.T.Monteiro, J.F.Guerreiro, J.P.Gonçalves, N.P.Mira, S.C.dos Santos, T.R.Cabrito, M.Palma, C.Costa, A.P.Francisco, S.C.Madeira, A.L.Oliveira, A.T.Freitas and I.Sá-Correia	D161–D166
S	**OnTheFly: a database of *Drosophila melanogaster* transcription factors and their binding sites**	S.Shazman, H.Lee, Y.Socol, R.S.Mann and B.Honig	D167–D171

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

S	**BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells**	D.Chacon, D.Beck, D.Perera, J.W.H.Wong and J.E.Pimanda	D172–D177
S	**DPRP: a database of phenotype-specific regulatory programs derived from transcription factor binding data**	D.T.W.Tzeng, Y.-T.Tseng, M.Ung, I.-E.Liao, C.-C.Liu and C.Cheng	D178–D183
S	**Network portal: a database for storage, analysis and visualization of biological networks**	S.Turkarslan, E.J.Wurtmann, W.-J.Wu, N.Jiang, J.C.Bare, K.Foley, D.J.Reiss, P.Novichkov and N.S.Baliga	D184–D190
	II. Protein sequence and structure, motifs and domains		
	Activities at the Universal Protein Resource (UniProt)	The UniProt Consortium	D191–D198
	Data, information, knowledge and principle: back to metabolism in KEGG	M.Kanehisa, S.Goto, Y.Sato, M.Kawashima, M.Furumichi and M.Tanabe	D199–D205
	The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)	R.Overbeek, R.Olson, G.D.Pusch, G.J.Olsen, J.J.Davis, T.Disz, R.A.Edwards, S.Gerdes, B.Parrello, M.Shukla, V.Vonstein, A.R.Wattam, F.Xia and R.Stevens	D206–D214
	UniCarbKB: building a knowledge platform for glycoproteomics	M.P.Campbell, R.Peterson, J.Mariethoz, E.Gasteiger, Y.Akune, K.F.Aoki-Kinoshita, F.Lisacek and N.H.Packer	D215–D221
	Pfam: the protein families database	R.D.Finn, A.Bateman, J.Clements, P.Coggill, R.Y.Eberhardt, S.R.Eddy, A.Heger, K.Hetherington, L.Holm, J.Mistry, E.L.L.Sonnhammer, J.Tate and M.Punta	D222–D230
S	**eggNOG v4.0: nested orthology inference across 3686 organisms**	S.Powell, K.Forslund, D.Szklarczyk, K.Trachana, A.Roth, J.Huerta-Cepas, T.Gabaldón, T.Rattei, C.Creevey, M.Kuhn, L.J.Jensen, C.von Mering and P.Bork	D231–D239
	Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis	J.G.Lees, D.Lee, R.A.Studer, N.L.Dawson, I.Sillitoe, S.Das, C.Yeats, B.H.Dessailly, R.Rentzsch and C.A.Orengo	D240–D245
S	**LenVarDB: database of length-variant protein domains**	E.Mutt, O.K.Mathew and R.Sowdhamini	D246–D250
	The Transporter Classification Database	M.H.Saier Jr, V.S.Reddy, D.G.Tamang and Å.Västermark	D251–D258
	The eukaryotic linear motif resource ELM: 10 years and counting	H.Dinkel, K.Van Roey, S.Michael, N.E.Davey, R.J.Weatheritt, D.Born, T.Speck, D.Krüger, G.Grebnev, M.Kubań, M.Strumillo, B.Uyar, A.Budd, B.Altenberg, M.Seiler, L.B.Chemes, J.Glavina, I.E.Sánchez, F.Diella and T.J.Gibson	D259–D266
	FireDB: a compendium of biological and pharmacologically relevant ligands	P.Maietta, G.Lopez, A.Carro, B.J.Pingilley, L.G.Leon, A.Valencia and M.L.Tress	D267–D272
	HRaP: database of occurrence of HomoRepeats and patterns in proteomes	M.Y.Lobanov, I.V.Sokolovskiy and O.V.Galzitskaya	D273–D278
	SIMAP—the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage	R.Arnold, F.Goldenberg, H.-W.Mewes and T.Rattei	D279–D284

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	PDBe: Protein Data Bank in Europe	A.Gutmanas, Y.Alhroub, G.M.Battle, J.M.Berrisford, E.Bochet, M.J.Conroy, J.M.Dana, M.A.Fernandez Montecelo, G.van Ginkel, S.P.Gore, P.Haslam, R.Hatherley, P.M.S.Hendrickx, M.Hirshberg, I.Lagerstedt, S.Mir, A.Mukhopadhyay, T.J.Oldfield, A.Patwardhan, L.Rinaldi, G.Sahni, E.Sanz-García, S.Sen, R.A.Slowley, S.Velankar, M.E.Wainwright and G.J.Kleywegt	D285–D291
	PDBsum additions	T.A.P.de Beer, K.Berka, J.M.Thornton and R.A.Laskowski	D292–D296
	MMDB and VAST+: tracking structural similarities between macromolecular complexes	T.Madej, C.J.Lanczycki, D.Zhang, P.A.Thiessen, R.C.Geer, A.Marchler-Bauer and S.H.Bryant	D297–D303
	SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures	N.K.Fox, S.E.Brenner and J.-M.Chandonia	D304–D309
	SCOP2 prototype: a new approach to protein structure mining	A.Andreeva, D.Howorth, C.Chothia, E.Kulesha and A.G.Murzin	D310–D314
S	**ArchDB 2014: structural classification of loops in proteins**	J.Bonet, J.Planas-Iglesias, J.Garcia-Garcia, M.A.Marín-López, N.Fernandez-Fuentes and B.Oliva	D315–D319
	IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners	S.Fukuchi, T.Amemiya, S.Sakamoto, Y.Nobe, K.Hosoda, Y.Kado, S.D.Murakami, R.Koike, H.Hiroaki and M.Ota	D320–D325
	pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins	M.Varadi, S.Kosol, P.Lebrun, E.Valentini, M.Blackledge, A.K.Dunker, I.C.Felli, J.D.Forman-Kay, R.W.Kriwacki, R.Pierattelli, J.Sussman, D.I.Svergun, V.N.Uversky, M.Vendruscolo, D.Wishart, P.E.Wright and P.Tompa	D326–D335
	ModBase, a database of annotated comparative protein structure models and associated resources	U.Pieper, B.M.Webb, G.Q.Dong, D.Schneidman-Duhovny, H.Fan, S.J.Kim, N.Khuri, Y.G.Spill, P.Weinkam, M.Hammel, J.A.Tainer, M.Nilges and A.Sali	D336–D346
	MP:PD—a data base of internal packing densities, internal packing defects and internal waters of helical membrane proteins	A.Rose, D.Theune, A.Goede and P.W.Hildebrand	D347–D351
	RepeatsDB: a database of tandem repeat protein structures	T.Di Domenico, E.Potenza, I.Walsh, R.Gonzalo Parra, M.Giollo, G.Minervini, D.Piovesan, A.Ihsan, C.Ferrari, A.V.Kajava and S.C.E.Tosatto	D352–D357
	The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases	S.Orchard, M.Ammari, B.Aranda, L.Breuza, L.Briganti, F.Broackes-Carter, N.H.Campbell, G.Chavali, C.Chen, N.del-Toro, M.Duesbury, M.Dumousseau, E.Galeota, U.Hinz, M.Iannuccelli, S.Jagannathan, R.Jimenez, J.Khadake, A.Lagreid, L.Licata, R.C.Lovering, B.Meldal, A.N.Melidoni, M.Milagros, D.Peluso, L.Perfetto, P.Porras, A.Raghunath, S.Ricard-Blum, B.Roechert, A.Stutz, M.Tognolli, K.van Roey, G.Cesareni and H.Hermjakob	D358–D363

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	iPfam: a database of protein family and domain interactions found in the Protein Data Bank	R.D.Finn, B.L.Miller, J.Clements and A.Bateman	D364–D373
	3did: a catalog of domain-based interactions of known three-dimensional structure	R.Mosca, A.Céol, A.Stein, R.Olivella and P.Aloy	D374–D379
	FunCoup 3.0: database of genome-wide functional coupling networks	T.Schmitt, C.Ogris and E.L.L.Sonnhammer	D380–D388
	KBDOCK 2013: a spatial classification of 3D protein domain family interactions	A.W.Ghoorah, M.-D.Devignes, M.Smaïl-Tabbone and D.W.Ritchie	D389–D395
	Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis	P.Blohm, G.Frishman, P.Smialowski, F.Goebels, B.Wachinger, A.Ruepp and D.Frishman	D396–D400
	STITCH 4: integration of protein–chemical interactions with user data	M.Kuhn, D.Szklarczyk, S.Pletscher-Frankild, T.H.Blicher, C.von Mering, L.J.Jensen and P.Bork	D401–D407
S	**UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks**	R.K.R.Kalathur, J.P.Pinto, M.A.Hernández-Prieto, R.S.R.Machado, D.Almeida, G.Chaurasia and M.E.Futschik	D408–D414
	Protein Ontology: a controlled structured network of protein entities	D.A.Natale, C.N.Arighi, J.A.Blake, C.J.Bult, K.R.Christie, J.Cowart, P.D'Eustachio, A.D.Diehl, H.J.Drabkin, O.Helfer, H.Huang, A.M.Masci, J.Ren, N.V.Roberts, K.Ross, A.Ruttenberg, V.Shamovsky, B.Smith, M.S.Yerramalla, J.Zhang, A.AlJanahi, I.Çelen, C.Gan, M.Lv, E.Schuster-Lezell and C.H.Wu	D415–D421
	GPCRDB: an information system for G protein-coupled receptors	V.Isberg, B.Vroling, R.van der Kant, K.Li, G.Vriend and D.Gloriam	D422–D425
	Laminin-database v.2.0: an update on laminins in health and neuromuscular disorders	D.C.F.Golbert, E.Santana-van-Vliet, A.S.Mundstein, V.Calfo, W.Savino and A.T.R.de Vasconcelos	D426–D429
	CentrosomeDB: a new generation of the centrosomal proteins database for *Human* and *Drosophila melanogaster*	J.M.C.Alves-Cruzeiro, R.Nogales-Cadenas and A.D.Pascual-Montano	D430–D436
	SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans	F.Romagné, D.Santesmasses, L.White, G.K.Sarangi, M.Mariotti, R.Hübler, A.Weihmann, G.Parra, V.N.Gladyshev, R.Guigó and S.Castellano	D437–D443
	Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides	A.Gautam, K.Chaudhary, S.Singh, A.Joshi, P.Anand, A.Tuknait, D.Mathur, G.C.Varshney and G.P.S.Raghava	D444–D449
	CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse	Q.Wang, J.Huang, H.Sun, J.Liu, J.Wang, Q.Wang, Q.Qin, S.Mei, C.Zhao, X.Yang, X.S.Liu and Y.Zhang	D450–D458
	III. Metabolic and signalling pathways, enzymes		
	The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/ Genome Databases	R.Caspi, T.Altman, R.Billington, K.Dreher, H.Foerster, C.A.Fulcher, T.A.Holland, I.M.Keseler, A.Kothari, A.Kubo, M.Krummenacker, M.Latendresse, L.A.Mueller, Q.Ong, S.Paley, P.Subhraveti, D.S.Weaver, D.Weerasinghe, P.Zhang and P.D.Karp	D459–D471

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	The Reactome pathway knowledgebase	D.Croft, A.F.Mundo, R.Haw, M.Milacic, J.Weiser, G.Wu, M.Caudy, P.Garapati, M.Gillespie, M.R.Kamdar, B.Jassal, S.Jupe, L.Matthews, B.May, S.Palatnik, K.Rothfels, V.Shamovsky, H.Song, M.Williams, E.Birney, H.Hermjakob, L.Stein and P.D'Eustachio	D472–D477
	SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database	T.Jewison, Y.Su, F.M.Disfany, Y.Liang, C.Knox, A.Maciejewski, J.Poelzer, J.Huynh, Y.Zhou, D.Arndt, Y.Djoumbou, Y.Liu, L.Deng, A.C.Guo, B.Han, A.Pon, M.Wilson, S.Rafatnia, P.Liu and D.S.Wishart	D478–D484
S	**The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes**	N.Furnham, G.L.Holliday, T.A.P.de Beer, J.O.B.Jacobsen, W.R.Pearson and J.M.Thornton	D485–D489
	The carbohydrate-active enzymes database (CAZy) in 2013	V.Lombard, H.Golaconda Ramulu, E.Drula, P.M.Coutinho and B.Henrissat	D490–D495
S	**EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases**	Y.Wang, Z.Liu, H.Cheng, T.Gao, Z.Pan, Q.Yang, A.Guo and Y.Xue	D496–D502
	***MEROPS*: the database of proteolytic enzymes, their substrates and inhibitors**	N.D.Rawlings, M.Waller, A.J.Barrett and A.Bateman	D503–D509
	ASD v2.0: updated content and novel features focusing on allosteric regulation	Z.Huang, L.Mou, Q.Shen, S.Lu, C.Li, X.Liu, G.Wang, S.Li, L.Geng, Y.Liu, J.Wu, G.Chen and J.Zhang	D510–D516
	MultitaskProtDB: a database of multitasking proteins	S.Hernández, G.Ferragut, I.Amela, J.Perez-Pons, J.Piñol, A.Mozo-Villarias, J.Cedano and E.Querol	D517–D520
	The Structure–Function Linkage Database	E.Akiva, S.Brown, D.E.Almonacid, A.E.Barber, A.F.Custer, M.A.Hicks, C.C.Huang, F.Lauck, S.T.Mashiyama, E.C.Meng, D.Mischel, J.H.Morris, S.Ojha, A.M.Schnoes, D.Stryke, J.M.Yunes, T.E.Ferrin, G.L.Holliday and P.C.Babbitt	D521–D530
S	**CPLM: a database of protein lysine modifications**	Z.Liu, Y.Wang, T.Gao, Z.Pan, H.Cheng, Q.Yang, Z.Cheng, A.Guo, J.Ren and Y.Xue	D531–D536
S	**topPTM: a new module of dbPTM for identifying functional post-translational modifications in transmembrane proteins**	M.-G.Su, K.-Y.Huang, C.-T.Lu, H.-J.Kao, Y.-H.Chang and T.-Y.Lee	D537–D545
	IV. Viruses, bacteria, protozoa and fungi		
	Ensembl Genomes 2013: scaling up access to genome-wide data	P.J.Kersey, J.E.Allen, M.Christensen, P.Davis, L.J.Falin, C.Grabmueller, D.S.T.Hughes, J.Humphrey, A.Kerhornou, J.Khobova, N.Langridge, M.D.McDowall, U.Maheswari, G.Maslen, M.Nuhn, C.K.Ong, M.Paulini, H.Pedro, I.Toneva, M.A.Tuli, B.Walts, G.Williams, D.Wilson, K.Youens-Clark, M.K.Monaco, J.Stein, X.Wei, D.Ware, D.M.Bolser, K.L.Howe, E.Kulesha, D.Lawson and D.M.Staines	D546–D552
	RefSeq microbial genomes database: new representation and annotation strategy	T.Tatusova, S.Ciufo, B.Fedorov, K.O'Neill and I.Tolstoy	D553–D559

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	IMG 4 version of the integrated microbial genomes comparative analysis system	V.M.Markowitz, I.-M.A.Chen, K.Palaniappan, K.Chu, E.Szeto, M.Pillay, A.Ratner, J.Huang, T.Woyke, M.Huntemann, I.Anderson, K.Billis, N.Varghese, K.Mavromatis, A.Pati, N.N.Ivanova and N.C.Kyrpides	D560–D567
	IMG/M 4 version of the integrated metagenome comparative analysis system	V.M.Markowitz, I.-M.A.Chen, K.Chu, E.Szeto, K.Palaniappan, M.Pillay, A.Ratner, J.Huang, I.Pagani, S.Tringe, M.Huntemann, K.Billis, N.Varghese, K.Tennessen, K.Mavromatis, A.Pati, N.N.Ivanova and N.C.Kyrpides	D568–D573
	DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements	H.Luo, Y.Lin, F.Gao, C.-T.Zhang and R.Zhang	D574–D580
	PATRIC, the bacterial bioinformatics database and analysis resource	A.R.Wattam, D.Abraham, O.Dalay, T.L.Disz, T.Driscoll, J.L.Gabbard, J.J.Gillespie, R.Gough, D.Hix, R.Kenyon, D.Machi, C.Mao, E.K.Nordberg, R.Olson, R.Overbeek, G.D.Pusch, M.Shukla, J.Schulman, R.L.Stevens, D.E.Sullivan, V.Vonstein, A.Warren, R.Will, M.J.C.Wilson, H.S.Yoo, C.Zhang, Y.Zhang and B.W.Sobral	D581–D591
	BacDive—the Bacterial Diversity Metadatabase	C.Söhngen, B.Bunk, A.Podstawka, D.Gleim and J.Overmann	D592–D599
	EBI metagenomics—a new resource for the analysis and archiving of metagenomic data	S.Hunter, M.Corbett, H.Denise, M.Fraser, A.Gonzalez-Beltran, C.Hunter, P.Jones, R.Leinonen, C.McAnulla, E.Maguire, J.Maslen, A.Mitchell, G.Nuka, A.Oisel, S.Pesseat, R.Radhakrishnan, P.Rocca-Serra, M.Scheremetjew, P.Sterk, D.Vaughan, G.Cochrane, D.Field and S.-A.Sansone	D600–D606
S	**The Global Genome Biodiversity Network (GGBN) Data Portal**	G.Droege, K.Barker, J.J.Astrin, P.Bartels, C.Butler, D.Cantrill, J.Coddington, F.Forest, B.Gemeinholzer, D.Hobern, J.Mackenzie-Dodds, É.Ó Tuama, G.Petersen, O.Sanjur, D.Schindel and O.Seberg	D607–D612
	LPSN—list of prokaryotic names with standing in nomenclature	A.C.Parte	D613–D616
S	**MetaRef: a pan-genomic database for comparative and community microbial genomics**	K.Huang, A.Brady, A.Mahurkar, O.White, D.Gevers, C.Huttenhower and N.Segata	D617–D624
S	**POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes**	Y.Lan, J.C.Morrison, R.Hershberg and G.L.Rosen	D625–D632
S	**Ribosomal Database Project: data and tools for high throughput rRNA analysis**	J.R.Cole, Q.Wang, J.A.Fish, B.Chai, D.M.McGarrell, Y.Sun, C.T.Brown, A.Porras-Alfaro, C.R.Kuske and J.M.Tiedje	D633–D642
S	**The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks**	P.Yilmaz, L.W.Parfrey, P.Yarza, J.Gerken, E.Pruesse, C.Quast, T.Schweer, J.Peplies, W.Ludwig and F.O.Glöckner	D643–D648
S	**COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia**	P.Meysman, P.Sonego, L.Bianco, Q.Fu, D.Ledezma-Tejeida, S.Gama-Castro, V.Liebens, J.Michiels, K.Laukens, K.Marchal, J.Collado-Vides and K.Engelen	D649–D653

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	DOOR 2.0: presenting operons and their functions through dynamic and integrated views	X.Mao, Q.Ma, C.Zhou, X.Chen, H.Zhang, J.Yang, F.Mao, W.Lai and Y.Xu	D654–D659
	Virus Variation Resource—recent updates and future directions	J.R.Brister, Y.Bao, S.A.Zhdanov, Y.Ostapchuck, V.Chetvernin, B.Kiryutin, L.Zaslavsky, M.Kimelman and T.A.Tatusova	D660–D665
S	**CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes**	T.Fujisawa, S.Okamoto, T.Katayama, M.Nakao, H.Yoshimura, H.Kajiya-Kanegae, S.Yamamoto, C.Yano, Y.Yanaka, H.Maita, T.Kaneko, S.Tabata and Y.Nakamura	D666–D670
S	**HoPaCI-DB: host-*Pseudomonas* and *Coxiella* interaction database**	S.Bleves, I.Dunger, M.C.Walter, D.Frangoulidis, G.Kastenmüller, R.Voulhoux and A.Ruepp	D671–D676
	PortEco: a resource for exploring bacterial biology through high-throughput data and analysis tools	J.C.Hu, G.Sherlock, D.A.Siegele, S.A.Aleksander, C.A.Ball, J.Demeter, S.Gouni, T.A.Holland, P.D.Karp, J.E.Lewis, N.M.Liles, B.K.McIntosh, H.Mi, A.Muruganujan, F.Wymore and P.D.Thomas	D677–D684
S	**SporeWeb: an interactive journey through the complete sporulation cycle of *Bacillus subtilis***	R.T.Eijlander, A.de Jong, A.O.Krawczyk, S.Holsappel and O.P.Kuipers	D685–D691
	***Subti*Wiki–a database for the model organism *Bacillus subtilis* that links pathway, interaction and expression information**	R.H.Michna, F.M.Commichau, D.Tödter, C.P.Zschiedrich and J.Stülke	D692–D698
	MycoCosm portal: gearing up for 1000 fungal genomes	I.V.Grigoriev, R.Nikitin, S.Haridas, A.Kuo, R.Ohm, R.Otillar, R.Riley, A.Salamov, X.Zhao, F.Korzeniewski, T.Smirnova, H.Nordberg, I.Dubchak and I.Shabalov	D699–D704
	The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations	G.C.Cerqueira, M.B.Arnaud, D.O.Inglis, M.S.Skrzypek, G.Binkley, M.Simison, S.R.Miyasato, J.Binkley, J.Orvis, P.Shah, F.Wymore, G.Sherlock and J.R.Wortman	D705–D710
	The *Candida* Genome Database: The new homology information page highlights protein similarity and phylogeny	J.Binkley, M.B.Arnaud, D.O.Inglis, M.S.Skrzypek, P.Shah, F.Wymore, G.Binkley, S.R.Miyasato, M.Simison and G.Sherlock	D711–D716
	***Saccharomyces* genome database provides new regulation data**	M.C.Costanzo, S.R.Engel, E.D.Wong, P.Lloyd, K.Karra, E.T.Chan, S.Weng, K.M.Paskov, G.R.Roe, G.Binkley, B.C.Hitz and J.M.Cherry	D717–D725
	LoQAtE—Localization and Quantitation ATlas of the yeast proteomE. A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast	M.Breker, M.Gymrek, O.Moldavski and M.Schuldiner	D726–D730
S	**YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae***	H.Kim, J.Shin, E.Kim, H.Kim, S.Hwang, J.E.Shim and I.Lee	D731–D736
S	**BacMet: antibacterial biocide and metal resistance genes database**	C.Pal, J.Bengtsson-Palme, C.Rensing, E.Kristiansson and D.G.J.Larsson	D737–D743
	mVOC: a database of microbial volatiles	M.C.Lemfack, J.Nickel, M.Dunkel, R.Preissner and B.Piechulla	D744–D748

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

V. Human genome, model organisms, comparative genomics

	Ensembl 2014	P.Flicek, M.R.Amode, D.Barrell, K.Beal, K.Billis, S.Brent, D.Carvalho-Silva, P.Clapham, G.Coates, S.Fitzgerald, L.Gil, C.G.Girón, L.Gordon, T.Hourlier, S.Hunt, N.Johnson, T.Juettemann, A.K.Kähäri, S.Keenan, E.Kulesha, F.J.Martin, T.Maurel, W.M.McLaren, D.N.Murphy, R.Nag, B.Overduin, M.Pignatelli, B.Pritchard, E.Pritchard, H.S.Riat, M.Ruffier, D.Sheppard, K.Taylor, A.Thormann, S.J.Trevanion, A.Vullo, S.P.Wilder, M.Wilson, A.Zadissa, B.L.Aken, E.Birney, F.Cunningham, J.Harrow, J.Herrero, T.J.P.Hubbard, R.Kinsella, M.Muffato, A.Parker, G.Spudich, A.Yates, D.R.Zerbino and S.M.J.Searle	D749–D755
	RefSeq: an update on mammalian reference sequences	K.D.Pruitt, G.R.Brown, S.M.Hiatt, F.Thibaud-Nissen, A.Astashyn, O.Ermolaeva, C.M.Farrell, J.Hart, M.J.Landrum, K.M.McGarvey, M.R.Murphy, N.A.O'Leary, S.Pujar, B.Rajput, S.H.Rangwala, L.D.Riddick, A.Shkeda, H.Sun, P.Tamez, R.E.Tully, C.Wallin, D.Webb, J.Weber, W.Wu, M.DiCuccio, P.Kitts, D.R.Maglott, T.D.Murphy and J.M.Ostell	D756–D763
	The UCSC Genome Browser database: 2014 update	D.Karolchik, G.P.Barber, J.Casper, H.Clawson, M.S.Cline, M.Diekhans, T.R.Dreszer, P.A.Fujita, L.Guruvadoo, M.Haeussler, R.A.Harte, S.Heitner, A.S.Hinrichs, K.Learned, B.T.Lee, C.H.Li, B.J.Raney, B.Rhead, K.R.Rosenbloom, C.A.Sloan, M.L.Speir, A.S.Zweig, D.Haussler, R.M.Kuhn and W.J.Kent	D764–D770
	The Vertebrate Genome Annotation browser 10 years on	J.L.Harrow, C.A.Steward, A.Frankish, J.G.Gilbert, J.M.Gonzalez, J.E.Loveland, J.Mudge, D.Sheppard, M.Thomas, S.Trevanion and L.G.Wilming	D771–D779
	FlyBase 102—advanced approaches to interrogating FlyBase	S.E.St. Pierre, L.Ponting, R.Stefancsik, P.McQuilton and the FlyBase Consortium	D780–D788
	WormBase 2014: new views of curated biology	T.W.Harris, J.Baran, T.Bieri, A.Cabunoc, J.Chan, W.J.Chen, P.Davis, J.Done, C.Grove, K.Howe, R.Kishore, R.Lee, Y.Li, H.-M.Muller, C.Nakamura, P.Ozersky, M.Paulini, D.Raciti, G.Schindelman, M.A.Tuli, K.V.Auken, D.Wang, X.Wang, G.Williams, J.D.Wong, K.Yook, T.Schedl, J.Hodgkin, M.Berriman, P.Kersey, J.Spieth, L.Stein and P.W.Sternberg	D789–D793
S	**WormQTL^{HD}—a web database for linking human disease to natural variation data in *C. elegans***	K.J.van der Velde, M.de Haan, K.Zych, D.Arends, L.B.Snoek, J.E.Kammenga, R.C.Jansen, M.A.Swertz and Y.Li	D794–D801

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data	G.Koscielny, G.Yaikhom, V.Iyer, T.F.Meehan, H.Morgan, J.Atienza-Herrero, A.Blake, C.-K.Chen, R.Easty, A.Di Fenza, T.Fiegel, M.Grifiths, A.Horne, N.A.Karp, N.Kurbatova, J.C.Mason, P.Matthews, D.J.Oakley, A.Qazi, J.Regnart, A.Retha, L.A.Santos, D.J.Sneddon, J.Warren, H.Westerberg, R.J.Wilson, D.G.Melvin, D.Smedley, S.D.M.Brown, P.Flicek, W.C.Skarnes, A.-M.Mallon and H.Parkinson	D802–D809
	The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse	J.A.Blake, C.J.Bult, J.T.Eppig, J.A.Kadin, J.E.Richardson and The Mouse Genome Database Group	D810–D817
	The mouse Gene Expression Database (GXD): 2014 update	C.M.Smith, J.H.Finger, T.F.Hayamizu, I.J.McCright, J.Xu, J.Berghout, J.Campbell, L.E.Corbani, K.L.Forthofer, P.J.Frost, D.Miers, D.R.Shaw, K.R.Stone, J.T.Eppig, J.A.Kadin, J.E.Richardson and M.Ringwald	D818–D824
	Mouse Phenome Database	S.C.Grubb, C.J.Bult and M.A.Bogue	D825–D834
	EMAGE mouse embryo spatial gene expression database: 2014 update	L.Richardson, S.Venkataraman, P.Stevenson, Y.Yang, J.Moss, L.Graham, N.Burton, B.Hill, J.Rao, R.A.Baldock and C.Armit	D835–D844
	TISdb: a database for alternative translation initiation in mammalian cells	J.Wan and S.-B.Qian	D845–D850
S	**GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments**	F.Halbritter, A.I.Kousa and S.R.Tomlinson	D851–D858
	GWIPS-viz: development of a ribo-seq genome browser	A.M.Michel, G.Fox, A.M.Kiran, C.De Bo, P.B.F.O'Connor, S.M.Heaphy, J.P.A.Mullan, C.A.Donohue, D.G.Higgins and P.V.Baranov	D859–D864
	Current status and new features of the Consensus Coding Sequence database	C.M.Farrell, N.A.O'Leary, R.A.Harte, J.E.Loveland, L.G.Wilming, C.Wallin, M.Diekhans, D.Barrell, S.M.J.Searle, B.Aken, S.M.Hiatt, A.Frankish, M.-M.Suner, B.Rajput, C.A.Steward, G.R.Brown, R.Bennett, M.Murphy, W.Wu, M.P.Kay, J.Hart, J.Rajan, J.Weber, C.Snow, L.D.Riddick, T.Hunt, D.Webb, M.Thomas, P.Tamez, S.H.Rangwala, K.M.McGarvey, S.Pujar, A.Shkeda, J.M.Mudge, J.M.Gonzalez, J.G.R.Gilbert, S.J.Trevanion, R.Baertsch, J.L.Harrow, T.Hubbard, J.M.Ostell, D.Haussler and K.D.Pruitt	D865–D872
S	**Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants**	J.A.L.MacArthur, J.Morales, R.E.Tully, A.Astashyn, L.Gil, E.A.Bruford, P.Larsson, P.Flicek, R.Dalgleish, D.R.Maglott and F.Cunningham	D873–D878
	ExoLocator—an online view into genetic makeup of vertebrate proteins	A.A.Khoo, M.Ogrizek-Tomaš, A.Bulović, M.Korpar, E.Gürler, I.Slijepčević, M.Šikić and I.Mihalek	D879–D881

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

S	**Manteia, a predictive data mining system for vertebrate genes and its applications to human genetic diseases**	O.Tassy and O.Pourquié	D882–D891
S	**AgeFactDB—the JenAge Ageing Factor Database—towards data integration in ageing research**	R.Hühne, T.Thalheim and J.Sühnel	D892–D896
	PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome	J.Huerta-Cepas, S.Capella-Gutiérrez, L.P.Pryszcz, M.Marcet-Houben and T.Gabaldón	D897–D902
S	**1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans**	M.Pybus, G.M.Dall'Olio, P.Luisi, M.Uzkudun, A.Carreño-Torres, P.Pavlidis, H.Laayouni, J.Bertranpetit and J.Engelken	D903–D909
S	**dbPSHP: a database of recent positive selection across human populations**	M.J.Li, L.Y.Wang, Z.Xia, M.P.Wong, P.C.Sham and J.Wang	D910–D916
	Selectome update: quality control and computational improvements to a database of positive selection	S.Moretti, B.Laurenczy, W.H.Gharib, B.Castella, A.Kuzniar, H.Schabauer, R.A.Studer, M.Valle, N.Salamin, H.Stockinger and M.Robinson-Rechavi	D917–D921
	TreeFam v9: a new website, more species and orthology-on-the-fly	F.Schreiber, M.Patricio, M.Muffato, M.Pignatelli and A.Bateman	D922–D925
S	**Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments**	R.Petryszak, T.Burdett, B.Fiorelli, N.A.Fonseca, M.Gonzalez-Porta, E.Hastings, W.Huber, S.Jupp, M.Keays, N.Kryvych, J.McMurry, J.C.Marioni, J.Malone, K.Megy, G.Rustici, A.Y.Tang, J.Taubert, E.Williams, O.Mannion, H.E.Parkinson and A.Brazma	D926–D932
	GEISHA: an evolving gene expression resource for the chicken embryo	P.B.Antin, T.A.Yatskievych, S.Davey and D.K.Darnell	D933–D937
S	**The Gene Expression Barcode 3.0: improved data processing and mining tools**	M.N.McCall, H.A.Jaffee, S.J.Zelisko, N.Sinha, G.Hooiveld, R.A.Irizarry and M.J.Zilliox	D938–D943
S	**The pancreatic expression database: recent extensions and updates**	A.Z.Dayem Ullah, R.J.Cutts, M.Ghetia, E.Gadaleta, S.A.Hahn, T.Crnogorac-Jurcevic, N.R.Lemoine and C.Chelala	D944–D949
S	**CellFinder: a cell data repository**	H.Stachelscheid, S.Seltmann, F.Lekschas, J.-F.Fontaine, N.Mah, M.Neves, M.A.Andrade-Navarro, U.Leser and A.Kurtz	D950–D958
	Plasma Proteome Database as a resource for proteomics research: 2014 update	V.Nanjappa, J.K.Thomas, A.Marimuthu, B.Muthusamy, A.Radhakrishnan, R.Sharma, A.Ahmad Khan, L.Balakrishnan, N.A.Sahasrabuddhe, S.Kumar, B.N.Jhaveri, K.V.Sheth, R.Kumar Khatana, P.G.Shaw, S.M.Srikanth, P.P.Mathur, S.Shankar, D.Nagaraja, R.Christopher, S.Mathivanan, R.Raju, R.Sirdeshmukh, A.Chatterjee, R.J.Simpson, H.C.Harsha, A.Pandey and T.S.K.Prasad	D959–D965

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

VI. Genomic variation, diseases and drugs

	The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data	S.Köhler, S.C.Doelken, C.J.Mungall, S.Bauer, H.V.Firth, I.Bailleul-Forestier, G.C.M.Black, D.L.Brown, M.Brudno, J.Campbell, D.R.FitzPatrick, J.T.Eppig, A.P.Jackson, K.Freson, M.Girdea, I.Helbig, J.A.Hurst, J.Jähn, L.G.Jackson, A.M.Kelly, D.H.Ledbetter, S.Mansour, C.L.Martin, C.Moss, A.Mumford, W.H.Ouwehand, S.-M.Park, E.R.Riggs, R.H.Scott, S.Sisodiya, S.V.Vooren, R.J.Wapner, A.O.M.Wilkie, C.F.Wright, A.T.Vulto-van Silfhout, N.de Leeuw, B.B.A.de Vries, N.L.Washingthon, C.L.Smith, M.Westerfield, P.Schofield, B.J.Ruef, G.V.Gkoutos, M.Haendel, D.Smedley, S.E.Lewis and P.N.Robinson	D966–D974
	NCBI's Database of Genotypes and Phenotypes: dbGaP	K.A.Tryka, L.Hao, A.Sturcke, Y.Jin, Z.Y.Wang, L.Ziyabari, M.Lee, N.Popova, N.Sharopova, M.Kimura and M.Feolo	D975–D979
	ClinVar: public archive of relationships among sequence variation and human phenotype	M.J.Landrum, J.M.Lee, G.R.Riley, W.Jang, W.S.Rubinstein, D.M.Church and D.R.Maglott	D980–D985
S	**The Database of Genomic Variants: a curated collection of structural variation in the human genome**	J.R.MacDonald, R.Ziman, R.K.C.Yuen, L.Feuk and S.W.Scherer	D986–D992
	DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation	E.Bragin, E.A.Chatzimichali, C.F.Wright, M.E.Hurles, H.V.Firth, A.P.Bevan and G.J.Swaminathan	D993–D1000
	The NHGRI GWAS Catalog, a curated resource of SNP-trait associations	D.Welter, J.MacArthur, J.Morales, T.Burdett, P.Hall, H.Junkins, A.Klemm, P.Flicek, T.Manolio, L.Hindorff and H.Parkinson	D1001–D1006
	Lynx: a database and knowledge extraction engine for integrative medicine	D.Sulakhe, S.Balasubramanian, B.Xie, B.Feng, A.Taylor, S.Wang, E.Berrocal, U.Dave, J.Xu, D.Börnigen, T.C.Gilliam and N.Maltsev	D1007–D1012
	NECTAR: a database of codon-centric missense variant annotations	S.Gong, J.S.Ware, R.Walsh and S.A.Cook	D1013–D1019
	Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies	P.Papadopoulos, E.Viennas, V.Gkantouna, C.Pavlidis, M.Bartsakoulia, Z.-M.Ioannou, I.Ratbi, A.Sefiani, J.Tsaknakis, K.Poulas, G.Tzimas and G.P.Patrinos	D1020–D1026
	InvFEST, a database integrating information of polymorphic inversions in the human genome	A.Martínez-Fundichely, S.Casillas, R.Egea, M.Ràmia, A.Barbadilla, L.Pantano, M.Puig and M.Cáceres	D1027–D1032
	rSNPBase: a database for curated regulatory SNPs	L.Guo, Y.Du, S.Chang, K.Zhang and J.Wang	D1033–D1039
	canSAR: updated cancer research and drug discovery knowledgebase	K.C.Bulusu, J.E.Tym, E.A.Coker, A.C.Schierz and B.Al-Lazikani	D1040–D1047
	DriverDB: an exome sequencing database for cancer driver gene identification	W.-C.Cheng, I.-F.Chung, C.-Y.Chen, H.-J.Sun, J.-J.Fen, W.-C.Tang, T.-Y.Chang, T.-T.Wong and H.-W.Wang	D1048–D1054

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

	Progenetix: 12 years of oncogenomic data curation	H.Cai, N.Kumar, N.Ai, S.Gupta, P.Rath and M.Baudis	D1055–D1062
	Updates of the HbVar database of human hemoglobin variants and thalassemia mutations	B.Giardine, J.Borg, E.Viennas, C.Pavlidis, K.Moradkhani, P.Joly, M.Bartsakoulia, C.Riemer, W.Miller, G.Tzimas, H.Wajcman, R.C.Hardison and G.P.Patrinos	D1063–D1069
	HMDD v2.0: a database for experimentally supported human microRNA and disease associations	Y.Li, C.Qiu, J.Tu, B.Geng, J.Yang, T.Jiang and Q.Cui	D1070–D1074
	PubChem BioAssay: 2014 update	Y.Wang, T.Suzek, J.Zhang, J.Wang, S.He, T.Cheng, B.A.Shoemaker, A.Gindulyte and S.H.Bryant	D1075–D1082
S	**The ChEMBL bioactivity database: an update**	A.P.Bento, A.Gaulton, A.Hersey, L.J.Bellis, J.Chambers, M.Davies, F.A.Krüger, Y.Light, L.Mak, S.McGlinchey, M.Nowotka, G.Papadatos, R.Santos and J.P.Overington	D1083–D1090
	DrugBank 4.0: shedding new light on drug metabolism	V.Law, C.Knox, Y.Djoumbou, T.Jewison, A.C.Guo, Y.Liu, A.Maciejewski, D.Arndt, M.Wilson, V.Neveu, A.Tang, G.Gabriel, C.Ly, S.Adamjee, Z.T.Dame, B.Han, Y.Zhou and D.S.Wishart	D1091–D1097
	The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands	A.J.Pawson, J.L.Sharman, H.E.Benson, E.Faccenda, S.P.H.Alexander, O.P.Buneman, A.P.Davenport, J.C.McGrath, J.A.Peters, C.Southan, M.Spedding, W.Yu, A.J.Harmar and NC-IUPHAR	D1098–D1106
	SuperPain—a resource on pain-relieving compounds targeting ion channels	B.O.Gohlke, R.Preissner and S.Preissner	D1107–D1112
	The Transformer database: biotransformation of xenobiotics	M.F.Hoffmann, S.C.Preissner, J.Nickel, M.Dunkel, R.Preissner and S.Preissner	D1113–D1117
	Therapeutic target database update 2014: a resource for targeted therapeutics	C.Qin, C.Zhang, F.Zhu, F.Xu, S.Y.Chen, P.Zhang, Y.H.Li, S.Y.Yang, Y.Q.Wei, L.Tao and Y.Z.Chen	D1118–D1123
	Updates on the web-based VIOLIN vaccine database and analysis system	Y.He, R.Racz, S.Sayers, Y.Lin, T.Todd, J.Hur, X.Li, M.Patel, B.Zhao, M.Chung, J.Ostrow, A.Sylora, P.Dungarani, G.Ulysse, K.Kochhar, B.Vidri, K.Strait, G.W.Jourdian and Z.Xiang	D1124–D1132
S	**bNAber: database of broadly neutralizing HIV antibodies**	A.M.Eroshkin, A.LeBlanc, D.Weekes, K.Post, Z.Li, A.Rajput, S.T.Butera, D.R.Burton and A.Godzik	D1133–D1139
	SAbDab: the structural antibody database	J.Dunbar, K.Krawczyk, J.Leem, T.Baker, A.Fuchs, G.Georges, J.Shi and C.M.Deane	D1140–D1146
S	**AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses**	A.Qureshi, N.Thakur, H.Tandon and M.Kumar	D1147–D1153
S	**CAMP: Collection of sequences and structures of antimicrobial peptides**	F.H.Waghu, L.Gopi, R.S.Barai, P.Ramteke, B.Nizami and S.Idicula-Thomas	D1154–D1158

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

VII. Plant databases

	Recent updates and developments to plant genome size databases	S.Garcia, I.J.Leitch, A.Anadon-Rosell, M.Á.Canela, F.Gálvez, T.Garnatje, A.Gras, O.Hidalgo, E.Johnston, G.Mas de Xaxars, J.Pellicer, S.Siljak-Yakovlev, J.Vallès, D.Vitales and M.D.Bennett	D1159–D1166
S	**GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology**	Ž.Ramšak, Š.Baebler, A.Rotter, M.Korbar, I.Mozetič, B.Usadel and K.Gruden	D1167–D1175
S	**P-MITE: a database for plant miniature inverted-repeat transposable elements**	J.Chen, Q.Hu, Y.Zhang, C.Lu and H.Kuang	D1176–D1181
S	**PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors**	J.Jin, H.Zhang, L.Kong, G.Gao and J.Luo	D1182–D1187
	ppdb: plant promoter database version 3.0	A.Hieno, H.A.Naznin, M.Hyakumachi, T.Sakurai, M.Tokizawa, H.Koyama, N.Sato, T.Nishiyama, M.Hasebe, A.D.Zimmer, D.Lang, R.Reski, S.A.Rensing, J.Obokata and Y.Y.Yamamoto	D1188–D1192
S	**Gramene 2013: comparative plant genomics resources**	M.K.Monaco, J.Stein, S.Naithani, S.Wei, P.Dharmawardhana, S.Kumari, V.Amarasinghe, K.Youens-Clark, J.Thomason, J.Preece, S.Pasternak, A.Olson, Y.Jiao, Z.Lu, D.Bolser, A.Kerhornou, D.Staines, B.Walts, G.Wu, P.D'Eustachio, R.Haw, D.Croft, P.J.Kersey, L.Stein, P.Jaiswal and D.Ware	D1193–D1199
	LSD 2.0: an update of the leaf senescence database	Z.Li, Y.Zhao, X.Liu, J.Peng, H.Guo and J.Luo	D1200–D1205
S	**P^{3}DB 3.0: From plant phosphorylation sites to protein networks**	Q.Yao, H.Ge, S.Wu, N.Zhang, W.Chen, C.Xu, J.Gao, J.J.Thelen and D.Xu	D1206–D1213
	Manually curated database of rice proteins	P.Gour, P.Garg, R.Jain, S.V.Joseph, A.K.Tyagi and S.Raghuvanshi	D1214–D1221
	RiceWiki: a wiki-based database for community curation of rice genes	Z.Zhang, J.Sang, L.Ma, G.Wu, H.Wu, D.Huang, D.Zou, S.Liu, A.Li, L.Hao, M.Tian, C.Xu, X.Wang, J.Wu, J.Xiao, L.Dai, L.-L.Chen, S.Hu and J.Yu	D1222–D1228
	CottonGen: a genomics, genetics and breeding database for cotton research	J.Yu, S.Jung, C.-H.Cheng, S.P.Ficklin, T.Lee, P.Zheng, D.Jones, R.G.Percy and D.Main	D1229–D1236
	The Genome Database for Rosaceae (GDR): year 10 update	S.Jung, S.P.Ficklin, T.Lee, C.-H.Cheng, A.Blenda, P.Zheng, J.Yu, A.Bombarely, I.Cho, S.Ru, K.Evans, C.Peace, A.G.Abbott, L.A.Mueller, M.A.Olmstead and D.Main	D1237–D1244
S	**Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding**	T.Joshi, M.R.Fitzpatrick, S.Chen, Y.Liu, H.Zhang, R.Z.Endacott, E.C.Gaudiello, G.Stacey, H.T.Nguyen and D.Xu	D1245–D1252

VIII. Other databases

	DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research	C.Y.Seiler, J.G.Park, A.Sharma, P.Hunter, P.Surapaneni, C.Sedillo, J.Field, R.Algar, A.Price, J.Steel, A.Throop, M.Fiacco and J.LaBaer	D1253–D1260

Continued

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

Contents (Continued)

Volume 42, Database issue, January 1, 2014

MitoBreak: the mitochondrial DNA breakpoints database	J.Damas, J.Carneiro, A.Amorim and F.Pereira	D1261–D1268
NeXO Web: the NeXO ontology database and visualization platform	J.Dutkowski, K.Ono, M.Kramer, M.Yu, D.Pratt, B.Demchak and T.Ideker	D1269–D1274

S, *Supplementary Material available at NAR Online*

L, *Linked articles*

F, *article featured at NAR online*

Downloaded from https://academic.oup.com/nar/issue/42/D1 by University of Wisconsin-Madison Libraries user on 09 January 2023

D32–D37 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 11 November 2013*
doi:10.1093/nar/gkt1030

GenBank

Dennis A. Benson, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received September 18, 2013; Accepted October 7, 2013

ABSTRACT

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for over 280 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and GenBank staff assign accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through the National Center for Biotechnology Information (NCBI) Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, begin at the NCBI home page: www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centers. The US Patent and Trademark Office also contributes sequences from issued patents. GenBank participates with the EMBL Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) (2) and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes GenBank data available at no cost over the Internet, through FTP and a wide range of Web-based retrieval and analysis services (4).

RECENT DEVELOPMENTS

Updates to the submission portal and related tools

NCBI continues to expand a unified submission portal that will ultimately provide a single access point for data submitters (submit.ncbi.nlm.nih.gov). Submitters can create accounts that track and display all their submissions and that facilitate communication with relevant NCBI staff. With respect to GenBank, the portal now supports submissions of complete microbial genomes along with WGS genomes. Submitters may continue to use standard GenBank submission tools (see below) for other GenBank submissions.

In addition to the submission portal itself, GenBank now provides some additional tools for submitters. For submitters of prokaryotic ribosomal RNA data, NCBI submission tools (see below) now perform internal checks for proper strandedness or for chimeric sequences. For submitters of bacterial genomes, NCBI provides a genome annotation pipeline described on a new prokaryotic annotation page (www.ncbi.nlm.nih.gov/genome/annotation_prok/) that also gives guidance for its use. Links to this page appear on the GenBank home page (www.ncbi.nlm.nih.gov/genbank/) and the Genome home page (www.ncbi.nlm.nih.gov/genome/).

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

Published by Oxford University Press 2013. This work is written by US Government employees and is in the public domain in the US.

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Catalog files

Beginning in release 194, GenBank discontinued the legacy 'index' files that previously accompanied the sequence data files of GenBank FTP releases. In their place, GenBank now provides a set of new 'release catalog' products that reside in a directory named 'catalog' in the FTP release. These new files are tab-delimited, and each row represents a single GenBank record and includes the accession.version, GI number, molecule type, sequence length, organism, taxonomy ID, division code and BioProject and BioSample accessions. Other sets of files associate GenBank records with gene symbols or PubMed IDs. These new files are described in more detail in the GenBank release notes (ftp.ncbi.nih.gov/genbank/gbrel.txt).

TPA assemblies

GenBank now accepts submissions of third party annotation (TPA) assemblies, which are typically assemblies constructed from existing collections of reads already in the Sequence Read Archive (SRA). Submitters are now required to supply data in binary sequence alignment/map (BAM) format that should specify the alignments of individual reads to submitted scaffolds, and these alignments may contain gaps. TPA assembly sequences will have the keyword 'TPA:assembly' and will have definition lines that begin with 'TPA_asm'. An older example of such an assembly is DAAA00000000, which contains links both to the WGS contigs (DAAA02nnnnnn) and the individual scaffolds (e.g. GK000001–GK000030).

Changes to indexing of bacterial strains

In the past, GenBank has assigned unique taxonomy IDs for each particular strain of bacteria for which there were submitted sequence data. Given the rapid increase in the diversity of bacterial strains in data being submitted, in the near future GenBank will no longer assign taxonomy IDs at the strain level. Bacterial strains that already have a unique taxonomy ID will retain them, but any submitted data for a new strain will not be given a unique taxonomy ID; rather, GenBank will assign the taxonomy ID of the species to these data. Because these new strains will also be entered in the BioSample database as unique records, BioSample will become the source of unique NCBI identifiers for individual bacteria strains and isolates.

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank assigns sequence records to various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 12 taxonomic divisions [BCT, environmental sample sequence (ENV), INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL and VRT] and 6 high-throughput divisions [EST, GSS, high-throughput cDNA (HTC), high-throughput genomic (HTG) and STS]. Finally, the PAT division contains records supplied by patent offices, the transcriptome shotgun assembly (TSA) division contains sequences from TSA projects and the WGS division contains sequences from WGS projects. The size and growth of these divisions, and GenBank as a whole, are shown in Table 1.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/taxonomy/) developed by NCBI in collaboration with EMBL-Bank and DDBJ and with the valuable assistance of external advisers and curators (5). Over 280 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ and EMBL-Bank). The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data itself are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. The initial version of a sequence has the extension ".1". In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a *GI* number that also appears on the VERSION line following the *Accession.version*:

```
ACCESSION AF000001
VERSION AF000001.5 GI: 7274584
```

When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers. The Revision History report, available from the 'Display Settings' menu on the sequence record view, summarizes the various updates for that GenBank record.

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. /protein_id = "AAF14809.1". Protein sequence translations also receive their own unique GI number, which appears as a second qualifier on the CDS feature:

```
/db_xref = "GI : 6513858"
```

Citing GenBank records

Besides being the primary identifier of a GenBank sequence record, GenBank accessions are also the most

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 197 (8/2013)	Annual increase (%)[a]
WGS	Whole-genome shotgun data	500 420 412 665	62.4
TSA	Transcriptome shotgun data	8 633 123 935	49.9
PHG	Phages	119 812 712	42.5
VRL	Viruses	1 757 202 472	22.9
BCT	Bacteria	10 281 048 518	21.8
ENV	Environmental samples	3 743 277 434	10.9
INV	Invertebrates	2 737 140 646	9.8
PAT	Patented sequences	13 290 161 247	9.7
PLN	Plants	5 963 882 822	8.8
GSS	Genome survey sequences	23 726 384 753	8.1
VRT	Other vertebrates	3 068 956 026	6.3
MAM	Other mammals	911 342 025	5.6
HTG	High-throughput genomic	25 184 819 955	3.4
HTC	High-throughput cDNA	656 196 063	2.7
UNA	Unannotated	130 510	2.1
EST	Expressed sequence tags	41 665 629 009	1.9
PRI	Primates	6 425 093 034	1.7
SYN	Synthetic	941 078 074	1.4
ROD	Rodents	4 451 315 297	0.4
STS	Sequence tagged sites	636 326 479	0.0
TOTAL	All GenBank sequences	654 613 333 676	45.1

[a]Measured relative to Release 191 (8/2012).

Table 2. Top organisms in GenBank (Release 197)

Organism	Non-WGS base pairs
Homo sapiens	17 111 514 261
Mus musculus	9 982 065 736
Rattus norvegicus	6 524 450 090
Bos taurus	5 389 432 575
Zea mays	5 071 648 554
Sus scrofa	4 889 229 566
Danio rerio	3 119 512 595
Hordeum vulgare	1 463 247 509
Strongylocentrotus purpuratus	1 435 237 072
Oryza sativa Japonica Group	1 263 872 842
Macaca mulatta	1 256 717 300
Xenopus (Silurana) tropicalis	1 249 741 450
Nicotiana tabacum	1 198 798 076
Arabidopsis thaliana	1 152 899 341
Triticum aestivum	1 139 790 400
Drosophila melanogaster	1 127 199 957
Vitis vinifera	1 069 944 084
Pan troglodytes	1 008 818 677
Solanum lycopersicum	966 234 744
Canis lupus familiaris	952 526 510

efficient and reliable way to cite a sequence record in publications. We certainly encourage submitters and other authors to cite GenBank data using these accessions. However, as discussed earlier, since searching with a GenBank accession number will retrieve the most recent version of a record, the data returned from such searches will change over time if the record is updated. It is quite possible, therefore, for the sequence data retrieved today by an accession to be different from that discussed or analyzed in a paper published several years ago. We therefore recommend that authors include the version suffix when citing a GenBank accession (e.g. AF000001.5), so that future readers can easily retrieve the data in question.

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL-Bank and DDBJ, are submitted either by individual authors to one of the three databases or by sequencing centers as batches of EST, STS, GSS, HTC, TSA, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL-Bank so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/genbank/), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of ~3500 per day. The accession number serves as confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program *tbl2asn*, described at www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html. Submitters can keep abreast of updates to *tbl2asn* and Sequin by subscribing to the NCBI submissions RSS feed (www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey=genbanksubmissiontoo).

Submission using BankIt

About a third of author submissions are received through an NCBI Web-based data submission tool named BankIt. Using BankIt, authors enter sequence information and biological annotations, such as coding regions or mRNA features, directly into a series of tabbed forms that allow the submitter to describe the sequence further without

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

having to learn formatting rules or controlled vocabularies. In addition, BankIt allows submitters to upload source and annotation data using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of BLAST called Vecscreen.

Submission using Sequin and tbl2asn

NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (such as a single cDNA), phylogenetic studies, population studies, mutation studies, environmental samples with or without alignments and sequences with complex annotation. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. In addition, Sequin is able to accommodate large sequences, such as the 5.6-Mb *Escherichia coli* genome, and read in a full complement of annotations from simple tables. The most recent version, Sequin 12.30, was released in November 2012 and is available for Macintosh, PC and Unix computers via anonymous FTP at ftp.ncbi.nlm.nih.gov/sequin. Once a submission is completed, submitters can e-mail the Sequin file to gb-sub@ncbi.nlm.nih.gov or upload the Sequin file to www.ncbi.nlm.nih.gov/LargeDirSubs/dir_submit.cgi. Submitters of large, heavily annotated genomes may find it convenient to use the command line tool *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank. These files for WGS genome and TSA submissions are then transmitted to GenBank via the Submission Portal.

Notes on particular divisions

Environmental sample sequences

The ENV division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an "/environmental_sample" qualifier in the source feature. Environmental sample sequences are generally submitted for whole metagenomic shotgun sequencing experiments or surveys of sequences from targeted genes, like 16S rRNA. NCBI continues to support BLAST searches (see below) of metagenomic ENV sequences, but sequences within WGS projects are now part of the WGS BLAST database.

Whole-genome shotgun sequences

WGS sequences appear in GenBank as groups of sequence-overlap contigs collected under a master WGS record. Each master record represents a WGS project and has an accession number in the Nucleotide database consisting of a four-letter prefix followed by eight zeroes and a version suffix as found in standard GenBank records. The number of zeroes increases to nine for WGS projects with 1 million or more contigs. Master records contain no sequence data; rather, links appear at the bottom of these records that provide displays of individual contigs in the WGS browser. Contig records have accessions consisting of the same four-letter prefix as their master accession, followed by a two-digit version number and a six-digit contig ID. For example, the WGS accession number 'AAAA02002744' is assigned to contig number '002744' of the second version of project 'AAAA', whose accession number is 'AAAA00000000.2'. Currently, there are over 12 000 WGS sequencing projects, many of whose data have been used to build over 14 million scaffolds and chromosomes for genome assemblies. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form "/experimental = *text*" and "/inference = *TYPE*: *text*", where *TYPE* is one of a number of standard inference types and *text* consists of structured text.

Transcriptome shotgun assembly sequences

The TSA division contains TSA sequences that are assembled from sequences deposited in the NCBI Trace Archive, the SRA and the EST division of GenBank. While neither the Trace Archive nor SRA is a part of GenBank, they are part of the INSDC and provide access to the data underlying these assemblies (4,6). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'. TSA continues to be one of the most rapidly growing divisions of GenBank, growing by 50% in the past year.

Genome survey sequences

GSS data continue to be an important resource for genome sequencing and annotation studies, and at almost 24 billion base pairs, the GSS division remains one of the largest non-WGS divisions in GenBank. GSS data are available for download from ftp.ncbi.nlm.nih.gov/repository/dbGSS/ and from the GenBank FTP site. In the near future GenBank will begin accepting GSS sequences with annotations provided by submitters. Because the data model for dbGSS does not allow for annotations, these new sequences will be indexed in the Entrez Nucleotide database (see below) and can be retrieved with the query 'gbdiv gss[prop]'.

High-throughput genomic and high-throughput cDNA sequences

The HTG division of GenBank (www.ncbi.nlm.nih.gov/genbank/htgs/) contains unfinished large-scale genomic

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

records, which are in transition to a finished state (7). These records are designated as belonging to Phases 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains HTC sequences that are of draft quality but may contain 5′ UTRs, 3′ UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in (8).

Special Record types

Third party annotation

TPA records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (www.ncbi.nlm.nih.gov/genbank/TPA). In addition to the new TPA assembly sequences discussed earlier, each of the 167 000 TPA records falls into one of three categories: 'experimental', in which case there is direct experimental evidence for the existence of the annotated molecule; 'inferential', in which case the experimental evidence is indirect; and 'assembly', where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label "TPA_exp:", "TPA_inf:" or "TPA_asm:" at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records

Within GenBank, CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is CM000663 for human chromosome 1.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (4). Records from the EST and GSS divisions of GenBank are stored in the EST and GSS databases, whereas all other GenBank records are stored in the Nucleotide database. GenBank sequences that are part of population or phylogenetic studies are also collected together in the PopSet database, and conceptual translations of CDS sequences annotated on GenBank records are available in the Protein database. Each of these databases is linked to the scientific literature in PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Education page (www.ncbi.nlm.nih.gov/education/).

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The BioProject database (www.ncbi.nlm.nih.gov/bioproject), developed at NCBI and subsequently adopted across the INSDC, allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. BioProject includes pointers to data from a wide variety of projects deposited in any NCBI primary data archive. Sequencing projects focus on genomes, metagenomes, transcriptomes, comparative genomics and on particular loci, such as 16S ribosomal RNA. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. In addition, sequence records may now have a link to the BioSample database (9) that provides additional information about the biological materials used in the study that produced the sequence data. Such studies include genome wide association studies, high-throughput sequencing, microarrays and epigenomic analyses. As an example, the TSA project GAAA contains DBLINK lines that associate the GenBank sequence record with BioProject record PRJNA54005 and BioSample record SRS283232, as well as the SRA record containing the raw data, SRR401852:

```
BioProject: PRJNA77699
BioSample: SRS283232
Sequence Read Archive: SRR401852
```

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (10,11). BLAST searches may be performed on the NCBI Web site (12) or by using a set of standalone programs distributed by FTP (4).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data

Case No.: IPR2023-00322
U.S. Patent No. 10,273,492

from EMBL-Bank and DDBJ, is available by anonymous FTP from NCBI at ftp.ncbi.nlm.nih.gov/genbank. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 197 there are 1905 files requiring 607 GB of uncompressed disk storage. A script is provided in ftp.ncbi.nlm.nih.gov/genbank/tools/ to convert a set of daily updates into a cumulative update.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 45, Room 6AN12D-37, 45 Center Drive, Bethesda, MD 20892, USA.

ELECTRONIC ADDRESSES

www.ncbi.nlm.nih.gov: NCBI Home Page.
gb-sub@ncbi.nlm.nih.gov: Submission of sequence data to GenBank.
update@ncbi.nlm.nih.gov: Revisions to, or notification of release of, 'confidential' GenBank entries.
info@ncbi.nlm.nih.gov: General information about NCBI resources.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
2. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeno-Tarraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
3. Ogasawara,O., Mashima,J., Kodama,Y., Kaminuma,E., Nakamura,Y., Okubo,K. and Takagi,T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.*, **41**, D25–D29.
4. NCBI Resource Coordinators. (2013) Database resources at the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
5. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
6. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
7. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA sequences to the databases. In: Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc, New York, NY, pp. 65–81.
8. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
12. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.