Communication

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201700139

Design Strategy of Multi-electron Transfer Catalysts Based on a Bioinformatic Analysis of Oxygen Evolution and Reduction Enzymes

Hideshi Ooka,[a, b] Kazuhito Hashimoto,[c] and Ryuhei Nakamura*[b, d]

Abstract: Understanding the design strategy of photosynthetic and respiratory enzymes is important to develop efficient artificial catalysts for oxygen evolution and reduction reactions. Here, based on a bioinformatic analysis of cyanobacterial oxygen evolution and reduction enzymes (photosystem II: PS II and cytochrome c oxidase: COX, respectively), the gene encoding the catalytic D1 subunit of PS II was found to be expressed individually across 38 phylogenetically diverse strains, which is in contrast to the operon structure of the genes encoding major COX subunits. Selective synthesis of the D1 subunit minimizes the repair cost of PS II, which allows compensation for its instability by lowering the turnover number required to generate a net positive energy yield. The different bioenergetics observed between PS II and COX suggest that in addition to the catalytic activity rationalized by the Sabatier principle, stability factors have also provided a major influence on the design strategy of biological multi-electron transfer enzymes.

Keywords: Bioenergetics · Bioinformatics · Catalysis · Life-cycle assessment · Oxygen evolution

There is currently an intensive interest in energy conversion systems such as fuel cells or solar-driven water electrolysis cells.[1] Multi-electron transfer reactions such as the oxygen evolution reaction (OER)[2,3] or the oxygen reduction reaction (ORR)[4,5] are indispensable components of these systems, and the development of efficient catalysts has been a central topic in the field of electrocatalysis. So far, *in-silico* methods such as the d-band theory have played a vital role in developing new catalysts.[6,7] Although scaling relationships are difficult to overcome experimentally,[8,9] the simplicity of binding energy optimization as a strategy, along with the rapid development of computational techniques, has yielded both active catalysts[10,11] and a deeper understanding of multi-electron transfer reactions in general.[6,12] Recent computational studies have addressed increasingly wider ranges of materials with the help of machine-learning techniques to reduce the computational cost.[13,14] Analysis of large experimental datasets has also shown to be possible using data mining approaches. For example, in a recent study by Shao-Horn and coworkers,[15] experimental data from various reports were collected and analyzed statistically after standardizing with the activity of a material common in all studies ($LaCoO_3$). This approach can be considered as a technical breakthrough in this field, because one of the difficulties hindering catalyst informatics was the lack of data in a uniform structure. The implementation of an internal standard allows access to the wealth of experimental data accumulated to this day, which may lead to a more widespread application of machine-learning techniques.

Here, we report our attempt to further advance *in-silico* catalyst design by developing a new framework to analyze the genetic information of multi-electron transfer enzymes. Extensive databases which contain the genetic sequence of

[a] *H. Ooka*
Department of Applied Chemistry,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[b] *H. Ooka, Dr. R. Nakamura*
Biofunctional Catalyst Research Team, RIKEN Center for Sustainable Resource Science (CSRS)
2-1 Hirosawa, Wako, Saitama 351-0198, Japan
phone: +81-(0)48-467-9539
fax: +81-(0)48-462-4639
E-mail: ryuhei.nakamura@riken.jp
[c] *Dr. K. Hashimoto*
National Institute for Materials Science (NIMS)
1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan
[d] *Dr. R. Nakamura*
Earth-Life Science Institute (ELSI)
Tokyo Institute of Technology,
2-12-1-lE-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.201700139

enzymes are readily available online. Furthermore, in contrast to experimentally-determined activity which deviates depending on the evaluation method (current vs overpotential, or pH conditions),[2,15] the unambiguity of the gene sequence allows for a more robust statistical analysis. Despite the enormous potential as a data resource however, no study has previously attempted to utilize genetic sequence information for further catalyst development.

Particularly in this study, we have attempted to identify how Nature addressed the life-cycle costs of OER (photosystem II, PS II) and ORR (cytochrome c oxidase, COX) enzymes based on an *in-silico* genetic analysis for the first time. This approach is in contrast to the extensively performed protein-level analysis which focuses on the molecular coordinates of atoms within the crystal structure. It is possible to analyze the structure-activity relationships of enzymes based on the d-band theory, in similar fashion to studies on inorganic bulk materials. Indeed, such information on the atomic spatial arrangement has yielded valuable insight into how enzymes achieve high activity.[16,17] However, the redundancy within the genetic code indicates that the same protein can be synthesized from a variety of different genetic sequences. This leads to a loss of information during DNA-to-protein conversion, and therefore, there is insight which can be uncovered only by genetic analysis.

In this study, we have evaluated the stability requirements of multi-electron transfer enzymes based on the genetic structures of the enzymes, because the energetics of enzyme repair are dictated by gene expression costs. The 38 strains of cyanobacteria (Table S1) used in this study are found in varying environments such as temperature, pH, and salinity, supporting the generality of statistically-relevant trends found within this dataset. Through this data-driven approach, we have attempted to uncover the difference in design criteria for the OER and ORR enzymes, which is often assumed to be dominated by activity (i.e. activation barrier) in the field of catalysis.

As shown in Figure 1A, PS II and COX catalyze OER and ORR.[1] Together, they compose a sustainable energy conversion cycle, much like how mankind may one day use solar-driven water electrolysis in combination with fuel cells.[1] The energy produced at PS II and COX are initially accumulated as a chemical potential gradient of protons which is later converted to ATP (adenosine triphosphate), the universal energy carrier in biology (Figure 1B). Eventually however, the multi-electron transfer enzymes will become deactivated, triggering a repair process (Figure 1C). The amount of ATP necessary to repair the enzyme is dictated by the length of the genetic sequence encoding the damaged region, and in terms of life-cycle assessment, the amount of ATP produced during OER / ORR (Figure 1B) must be greater than the amount of ATP required for enzyme repair (Figure 1C) to ensure a net positive energy gain. As stability is a critical requirement even for artificial systems,[18] we expect that seeking how Nature addressed

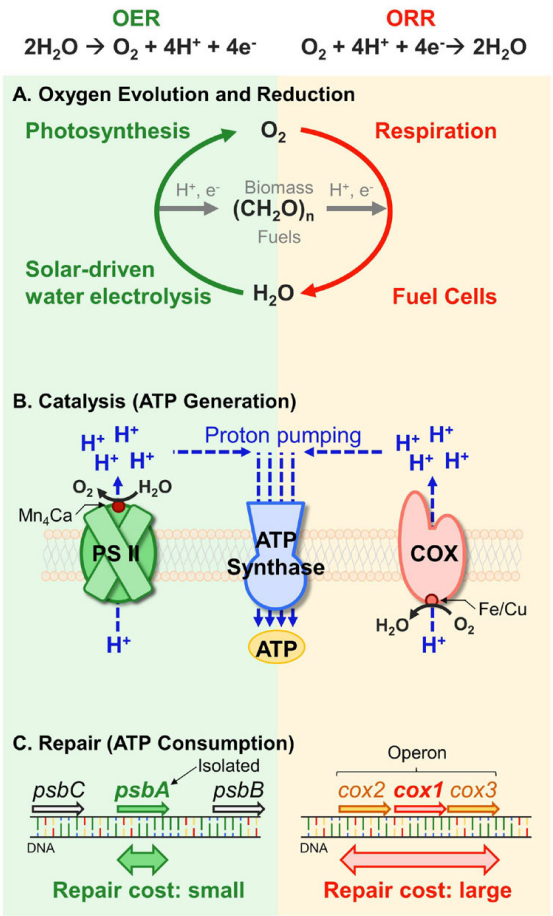

Figure 1. A: Schematic diagram of photosynthesis (enzyme: PS II), respiration (enzyme: COX), solar-driven water electrolysis, and fuel cells. B: PS II and COX generate energy (ATP) via OER and ORR. C: Upon enzyme deactivation, ATP is consumed to repair the catalyst. A longer genetic sequence increases the repair cost. The balance between ATP generation (stability) and consumption (repair) dictates the minimum turnover cycles necessary for a net positive ATP yield.

life-cycle costs of multi-electron transfer enzymes will provide a guideline for the development of artificial systems.[19]

In order to evaluate the repair cost of PS II and COX, their genetic structure was investigated. In the case of prokaryotes such as cyanobacteria, genetic information is often organized into structures called operons,[20] which allow the expression of nearby genes (<20 kilobasepairs, kbp) to be regulated simultaneously. Although an operon structure is beneficial because fewer regulatory regions would allow for smaller genome sizes, forming an operon would increase the repair cost due to the larger size of the coding region.

Figure 2A is a histogram which shows the operon structure of *psbA* and *cox1* compiled from the prokaryotic

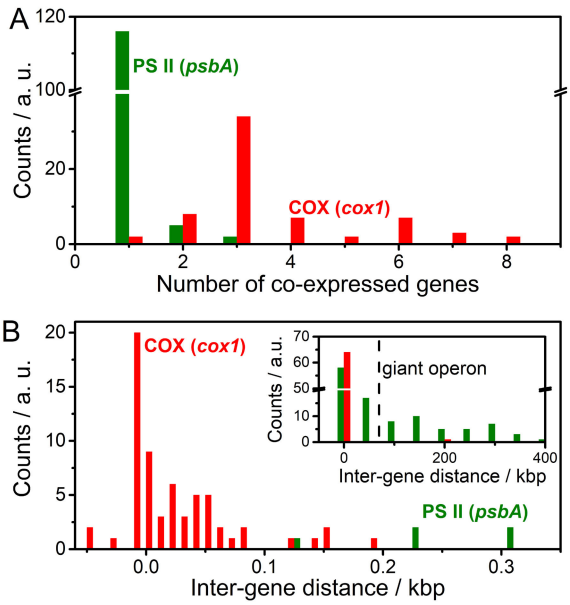

Figure 2. Genetic structure of *psbA* and *cox1* genes. A: Number of co-expressed genes based on the operon prediction of ProOpDB. (Y-axis break from 40 to 100) B: Distance between *psbA* and another PS II gene or *cox1* and another COX gene. Inset shows the inter-gene distance in a larger scale. The Y-axis in all panels show the number of genes within the investigated dataset which meets the criteria on the X-axis. The leftmost bar in the inset of Figure 2B shows 58 *psbA* genes with an inter-gene distance <50 kbp. However, the majority of them are not visible in the main panel because they are outside the scale limit of the X-axis (inter-gene distance >0.3 kbp).

operon database ProOpDB.[21,22] The Y-axis indicates the number of genes within the investigated dataset (123 *psbA* genes, 65 *cox1* genes) which are co-expressed with the number of genes indicated on the X-axis. The two genes encode the catalytic region of PS II and COX (*psbA*: D1 subunit of PS II, *cox1*: subunit I of COX), and therefore, biological optimization of OER and ORR catalysts are expected to be most pronounced in these two genes. According to this database, the majority of *psbA* genes (94%) were expressed individually, whereas *cox1* was almost always expressed together with two other COX genes (*cox2*, *cox3*).

Further support that *psbA* genes tend to be expressed individually compared to *cox1* can be found in Figure 2B. In the case of *cox1*, *cox2* and *cox3* are almost always less than 0.1 kbp away on the genome, indicating the catalytic site (*cox1*) is synthesized in concert with the substrate-binding site (*cox2*) and the transmembrane scaffold (*cox3*). On the other hand, the majority of *psbA* genes are outside of the scale in the main panel of Figure 2B. Upon increasing the scale of the X-axis (inset), it becomes apparent that the distance between *psbA* and its nearest PS II gene is several orders of magnitude larger, with many inter-gene distances which are larger than even the largest reported operons (70–80 kbp, dashed line in inset)[23] in literature. Therefore, although PS II in its entirety functions as a water oxidizing photo-anode, these results indicate the OER catalytic site (*psbA*) is synthesized separately from light-harvesting antennae complexes (*psbB*, *psbC*), other reaction center proteins (*psbD*), or secondary electron transport subunits (*psbE*, *psbF*), consistent with the conclusion of Figure 2A. We emphasize that the stark contrast in the genetic structure of OER and ORR enzymes could be uncovered only through a direct statistical comparison of PS II and COX. Although the separate regulation of the *psbA* gene is considered the norm based on biochemical observations of model organisms,[24,25] the statistical relevance presented here highlights the generality of genetic trends for the first time. Below, the genetic structure will be used to understand how biological enzymes addressed life-cycle costs and stability issues, which would yield an understanding of multi-electron transfer from a viewpoint outside the scope of the d-band theory.

To determine how the genetic structure of the active site influences the bioenergetics of PS II and COX, we estimated and compared the repair cost of each enzyme based on equation (1), where the first and second factors account for the transcription[26] and translation[27] costs, respectively.

$$\begin{aligned}\text{Repair Cost} &= 2\times(\text{DNA sequence length})\\ &+2\times(\text{protein sequence length})\ [\text{unit}:\ \text{ATP}]\end{aligned} \quad (1)$$

Here, sequence lengths refer to the amount of base pairs and amino acid monomers. The factor of two in the first term (transcription cost) is derived from the activation cost of a mono-phosphate nucleotide[26] whereas the factor of two in the second term (translation cost) is derived from the activation (1 GTP: guanosine triphosphate) and polymerization cost (1 ATP) per amino acid.[27]The enzyme repair costs calculated based on the gene expression costs of *psbA* and *cox1* using equation (1) are shown in Figure 3A. The selective replacement of the D1 subunit enabled by the individual expression of *psbA* genes benefits PS II significantly, as the repair cost of PS II is only one-third of that of COX.

As equation (1) does not consider post-translational events or synthesis of amino acid monomers which may require additional ATP, equation (1) represents the minimum repair cost of each enzyme. Therefore, a net positive ATP gain can be possible only if the following inequality is satisfied:

$$(\text{Repair Cost}) < (\text{TON}\times\text{ATP/cycle})\ [\text{ATP}] \quad (2)$$

Here, the term on the right represents the amount of ATP synthesized during the lifetime of the protein, where TON (turnover number) indicates the number of catalytic cycles performed prior to deactivation. Therefore, the

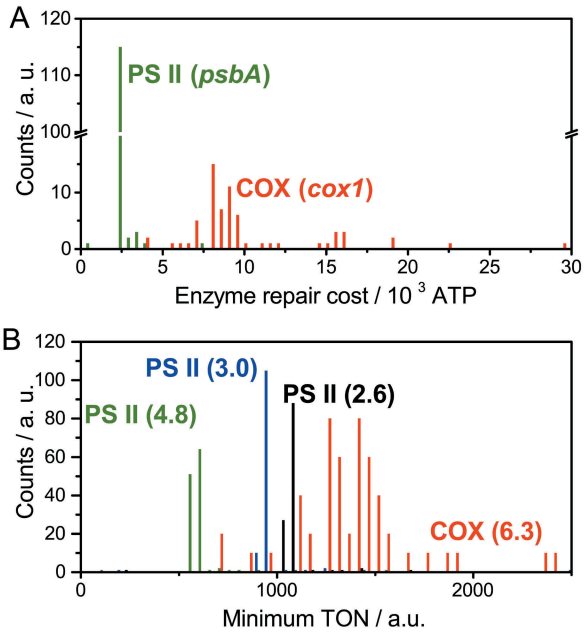

Figure 3. Histogram of the repair cost (A) and minimum TON (B) for PS II and COX. Numbers in parentheses indicate the ATP/cycle used for the calculation. There is a Y-axis break from 20 to 100 in panel A. The data count for COX in (B) has been amplified 10 fold due to the difference in scale with PS II.

minimum TON which satisfies equation (2) is expressed as (Repair cost) ÷ (ATP/cycle). As we have obtained the repair cost for each gene in Figure 3A, obtaining literature values for ATP/cycle allow calculation of the minimum TON.

Figure 3B shows the minimum TON for PS II and COX calculated using the literature values of ATP/cycle indicated in parentheses within the figure. While there is some discrepancy in the literature concerning the exact value of the ATP generated per OER cycle at PS II,[28] a stark contrast in the minimum TON between COX and PS II can be observed regardless of the exact number used.[28,29] As shown in Figure 3B, PS II requires only 500–1000 cycles to generate the ATP required for repair, because the OER catalytic site, encoded by the *psbA* gene, can be repaired separately from other PS II components such as the light-harvesting proteins. On the other hand, COX synthesizes all of its major components simultaneously, leading to an increased repair cost (Figure 3A). The larger cost can be covered only after 1500 ORR cycles have been performed. It should be noted that regulating a specific subunit individually is beneficial only if its repair takes place more frequently than the other subunits. Although other PS II components such as the D2, CP43, and PsbH are known to also require occasional repair,[30,31] the high rate of repair specific to the D1 subunit allows it to take full advantage of the minimized gene expression cost.[30,31]

The minimized repair cost of PS II, persistent through 38 strains of phylogenetically-diverse cyanobacteria, showcases contrasting bioenergetic strategies between OER and ORR enzymes. While deactivation of PS II under intense illumination is already well recognized,[28] the strategy of how Nature addressed the instability of PS II was revealed only by a direct genetic comparison of two energy-producing enzymes. The results of this study suggest instead of directly addressing this challenge by developing a stable OER catalyst, PS II chose to circumvent this problem by isolating the gene encoding the unstable D1 subunit (*psbA*) from other PS II genes (Figure 2). The separation of the unstable OER center with other PS II components such as light-harvesting pigments suppresses the repair cost of PS II sufficiently such that replacing D1 subunits upon being damaged becomes a viable strategy. As both stability and activity are major issues in artificial multi-electron catalysis, understanding how biological enzymes approached these challenges is important to realize active and robust catalysts. The method presented here supplements the activity-focused d-band theory with stability requirements based on the often-overlooked genomic structure, paving the way for a more comprehensive catalyst design encompassing both stability and activity requirements. As biomimetic and bioinspired approaches have guided material synthesis for decades, we expect catalyst development based on biological sequence information to become increasingly widespread in the near future.

Computational Methods

All cyanobacteria strains whose operon data were available within ProOpDB[22,23] as of November 2015 were included in this study. The complete list of strains and their copy number of *psbA* and *cox1* genes, as well as the basis of our calculations, are shown in the supplementary information.

Conflict of Interest

None declared.

Acknowledgements

The authors are grateful to Dr. Satoshi Kawaichi, Mr. Akira Murakami, and Mr. Akio Umezawa for help with the operon data analysis. This work was supported by JSPS Grant-in-Aid for Scientific Research no. 26288092. H.O. acknowledges Grant-in-Aid for JSPS Research Fellows no. 15J10450.

References

[1] H. Tributsch in *Carbon-Neutral Fuels and Energy Carriers* (Eds.: N. Z. Muradov, T. N. Veziroglu), CRC Press, Boca Raton, **2012**, pp. 415–463.
[2] H. Ooka, T. Takashima, A. Yamaguchi, T. Hayashi, R. Nakamura, *Chem. Commun.* **2017**, *53*, 7149–7161.
[3] O. Diaz-Morales, S. Raaijman, R. Kortlever, P. J. Kooyman, T. Wezendonk, J. Gascon, W. T. Fu, M. T. M. Koper, *Nat. Commun.* **2016**, *7*, 12363.
[4] V. R. Stamenkovic, B. Fowler, B. S. Mun, G. Wang, P. N. Ross, C. A. Lucas, N. M. Markovic, *Science* **2007**, *315*, 493–497.
[5] M. Nesselberger, S. Ashton, J. C. Meier, I. Katsounaros, K. J. J. Mayrhofer, M. Arenz, *J. Am. Chem. Soc.* **2011**, *133*, 17428–17433.
[6] J. K. Norskov, F. Studt, F. Abild-Pederson, T. Bligaard, *Fundamental Concepts in Heterogeneous Catalysis*, John Wiley and Sons, Hoboken, **2014**, pp. 114-137.
[7] F. Calle-Vallejo, J. I. Martinez, J. M. Garcia-Lastra, E. Abad, M. T. M. Koper, *Surf. Sci.* **2013**, *607*, 47–53.
[8] N. B. Halck, V. Petrykin, P. Krtil, J. Rossmeisl, *Phys. Chem. Chem. Phys.* **2014**, *16*, 13682–13688.
[9] X. Hong, K. Chan, C. Tsai, J. K. Norskov, *ACS Catal.* **2016**, *6*, 4428–4437.
[10] W. T. Hong, M. Risch, K. A . Stoerzinger, A. Grimaud, J. Suntivich, Y. Shao-Horn, *Energy Environ. Sci.* **2015**, *8*, 1404–1427.
[11] Z. W. She, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Norskov, T. F. Jaramillo, *Science* **2017**, *355*, eaad4998.
[12] A. J. Medford, A. Vojvodic, J. S. Hummelshoj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson, J. K. Norskov, *J. Catal.* **2015**, *328*, 36–42.
[13] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Norskov, *ACS Catal.* **2017**, *7*, 6600–6608.
[14] I. Takigawa, K. Shimizu, K. Tsuda, S. Takakusagi, *RSC Adv.* **2016**, *6*, 52587–52595.
[15] W. T. Hong, R. E. Welsch, Y. Shao-Horn, *J. Phys. Chem. C* **2016**, *120*, 78–86.
[16] J. Rossmeisl, K. Dimitrievski, P. Siegbahn, J. K. Norskov, *J. Phys. Chem. C* **2007**, *111*, 18821–18823.
[17] C. H. Kjaergaard, J. Rossmeisl, J. K. Norskov, *Inorg. Chem.* **2010**, *49*, 3567–3572.
[18] C. C. L. McCrory, S. Jung, I. M. Ferrer, S. M. Chatman, J. C. Peters, T. F. Jaramillo, *J. Am. Chem. Soc.* **2015**, *137*, 4347–4357.
[19] R. Sathre, C. D. Scown, W. R. Morrow, III, J. C. Stevens, I. D. Sharp, J. W. Ager, III, K. Walczak, F. A. Houle, J. B. Greenblatt, *Energy Environ. Sci.* **2014**, *7*, 3264–3278.
[20] S. Oehler, E. R. Eismann, H. Kramer, B. Muller-Hill, *EMBO J.* **1990**, *9*, 973–979.
[21] B. Taboada, R. Ciria, C. E. Martinez-Guerrero, E. Merino, *Nucleic Acids Res.* **2012**, *40*, D627–631.
[22] B. Taboada, C. Verde, E. Merino, *Nucleic Acids Res.* **2010**, *38*, e130.
[23] L. W. Hamoen, G. Venema, O. P. Kuipers, *Microbiology* **2003**, *149*, 9–17.
[24] P. Mulo, C. Sicora, E.-M. Aro, *Cell. Mol. Life Sci.* **2009**, *66*, 3697–3710.
[25] T. A. Nguyen, J. Brescic, D. J. Vinyard, T. Chandrasekar, G. C. Dismukes, *Mol. Biol. Evol.* **2012**, *29*, 35–38.
[26] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *THE CELL 5th Ed:* Newton Press, Boston, **2010** pp. 45–124.
[27] K. Moldave, *Annu. Rev. Biochem.* **1985**, *54*, 1109–1149.
[28] K. Miyata, K. Noguchi, I. Terashima, *Photosynth. Res.* **2012**, *113*, 165–180.
[29] M. Asashima *et al.*, *Life Science 3rd. Ed:* Yodosha, Tokyo, **2009** pp. 103.)
[30] M. A. K. Jansen, A. K. Mattoo, M. Edelman, *Eur. J. Biochem.* **1999**, *260*, 528–532.
[31] S. Jarvi, M. Suorsa, E.-M. Aro, *Biochim. Biophys. Acta* **2015**, *1847*, 900–909.

Received: November 13, 2017
Accepted: March 5, 2018
Published online on ■■, ■■■

